RISES ALL THE WAY UP: THE INTERPRETATION OF PROSODY, DISCOURSE ATTITUDES AND DIALOGUE STRUCTURE

Catherine Lai

A DISSERTATION

in

Linguistics

Presented to the Faculties of the University of Pennsylvania
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy

2012

Supervisor of Dissertation:

_____

Jiahong Yuan, Assistant Professor of Linguistics

Graduate Group Chair:

_____

Eugene Buckley, Associate Professor of Linguistics

Dissertation Committee:

Jiahong Yuan, Assistant Professor of Linguistics
Mark Liberman, Trustee Professor of Phonetics
Florian Schwarz, Assistant Professor of Linguistics

Acknowledgements

First of all, I would like to thank my advisor Jiahong Yuan. I came to Penn with a somewhat vague idea of doing something in the realm of semantics and pragmatics, but it didn't take long to get hooked on phonetics with Jiahong as a teacher. I've benefitted enormously from his mentoring through the years on several other projects in addition to dissertation work. The fact that I've been able to do a dissertation right on the intersection of semantics, pragmatics, and phonetics is a testament to his open mindedness, intellectual vigor and unfailing practical support.

Being at Penn has also allowed me the privilege of having Mark Liberman and Florian Schwarz on my dissertation committee. Mark never fails to amaze me with not only the breadth and depth of his knowledge on all areas of language research, but also his astounding clarity of thought. Mark's classes were critical for my developing the technical notions that have enabled me to investigate the prosody-meaning problem in a quantitative way. Florian's help was crucial in getting the semantic and pragmatic parts of this project on the right track. The way he has seen straight to the big conceptual problems through all the murky data I've thrown at him in the past years is truly impressive. Overall, I have to thank all of my committee for their continuous encouragement, vital criticism, and unbounded patience.

I'm extremely indebted to the participants of our regular phonetics lab speech lunch group for generous feedback on many occasions. The work presented here has been improved by feedback from many other people at various venues. I have had particularly helpful interactions with Yi Xu, Nigel Ward, Agustín Gravano, Yoon-sook Mo, Steve Isard, Sylvie Saget, Alan Black, Nick Campbell, Kristin Precoda, Andreas Stolcke, Maribel Romero, Chris Kennedy, Valentine Hacquard, participants at YRRSDS 2010/2011, but also many others! I also had the very good fortune of doing an internship at the Educational Testing Service. Although the work I did there was not directly related to my dissertation research, a lot of my thinking about prosody still owes a debt to my time there. So, special thanks to Klaus Zechner, Keelan Evanini, Su-Youn Yoon, and Derrick Higgins.

My dissertation wouldn't have been possible without the strong foundations I've gotten from the linguistic graduate program at Penn. I've learned a lot from all my teachers here, but I've been particularly inspired by Aravind Joshi, Scott Weinstein,

Charles Yang, Robin Clark, and Maribel Romero. I also need to thank Ani Nenkova for helpful feedback on my dissertation proposal. Very special thanks to Steven Bird at the University of Melbourne for steadfast support through my entire graduate career and for getting me interested in linguistics in the first place. Getting through the PhD would have been terrible without the help of Amy Forsyth, who has always been there to save the day from the administrative side and is just an all round great friend to have.

The keystone to all of this has really been my fellow graduate students. I need to give a special shout out the many p.lab companions I've had over the years: Marjorie Pak, Yanyan Sui, Jingjing Tan, Aviad 'Mr Sunshine' Eilam, Laia Mayol (P is for pragmatics too!), Josh Tauberer, Giang Nguyen, Yong-cheol Lee, Satoshi Nambu, and Keelan Evanini. We laughed, we cried, we gossipped ... we discussed everything from metrical structure to fashion, sometimes well past the witching hour.[1] This will always be the place I grew up academically. I'll miss you! I'd also like to give a special shout out to my honorary p.labites: Ariel Diertani, Laurel 'creaky' MacKenzie, Aaron Dinkin, Brittany McLaughlin, Caitlin Light, and Lucas Champollion for sharing their general fabulousness in so many ways. Many other people have made Penn a great place to be, so cheers to Lukasz Abramowicz, Dimka Atanassov, Jana Beck, Stefanie Brody, Toni Cook, Michael Friesner, Kyle Gorman, Lauren Friedman, Joe Fruehwald, Jonathan Gress-Right, Damien Hall, Robert Lannon, Marielle Lerner, Maya Ravindranath, Lydia Rieck, Neville Ryant, Tatjana Scheffler, Augustin Speyer, Meredith Tamminga, and Joel Wallenberg.

Sometimes it's easy to forget that life exists outside of the PhD, but my PhD definitely wouldn't exist without the great support of many outside the campus boundaries. I have the warmest thanks and love for the Tang family who have always looked after me as one of their own. My housemates at 540B made the initial transition to Philadelphia an easy one. Thanks to Nirav Mehta for a lot of laughs and for listening to a lot of raves and rants, and to Andrew Clausen for so much help over the years in many ways (but especially with the R!). But, of course, the Oscar goes to my parents, Christine and Kim Lai, for being the best, most patient and loving parents that anyone could possibly imagine. If I did anything good in my life, they get the credit. I promise to take you on that cruise!

[1] And we almost succeeded in getting to a mall in suburban Philadelphia on foot! Good times, crazy days.

ABSTRACT

RISES ALL THE WAY UP: THE INTERPRETATION OF PROSODY, DISCOURSE ATTITUDES AND DIALOGUE STRUCTURE

Catherine Lai

Supervisor: Jiahong Yuan

This dissertation is about what prosody contributes to dialogue interpretation. The view of prosody developed in this account is based on detailed quantitative investigations of the prosodic forms and interpretations of cue word and declarative responses, specifically with respect to the distribution and interpretation of terminal pitch rises. Drawing on results from corpus, production and perception studies, I argue that the underlying contribution of terminal rises is to signal that the dialogue has not come to a viable stopping point with respect to the task at hand. This approach enables us to explain previously incongruent findings about the connection between rises and attitudes like uncertainty. From this perspective, the perception of such attitudes does not arise directly from prosodic form, but instead depends upon a range of contextual factors. The experimental results indicate that the most important of these is how an utterance relates to the current question under discussion, rather than sentence or dialogue act type. However, variation in prosodic form is also affected by higher level factors like dialect, task, and speaker role: rises become more frequent on non-questioning moves as the need to co-ordinate becomes greater.

The experimental results allows us to make significant headway in clarifying the relationship between the prosodic, semantic and information structural properties of responses. This, in turn, sheds light on several outstanding questions about the con-

tribution of the rise in fall-rise accents and its relationship to information structural categories like contrastive topic. Overall, we see that rises don't act on the proposition that carries them, nor do they mark out specific IS categories. Instead they reveal the state of the discourse from the speaker's perspective. From a methodological point of view, I show that to gain a robust understanding the contribution of prosody on a particular meaning dimension, we need to take into account the baseline induced by the discourse configuration itself. These studies show the utility of using functional data analysis techniques to give more direct view of prosodic variation in larger datasets without manual prosodic annotation.

Contents

Acknowledgements ii

Abstract iv

Contents vi

List of Tables xi

List of Figures xiii

1 Introduction 1
- 1.1 Prosody in Dialogue . . . 1
- 1.2 The Research Problems and Background . . . 3
- 1.3 The Data . . . 7
- 1.4 Key Claims . . . 8
  - 1.4.1 Prosody, Discourse Attitude and Structure . . . 8
  - 1.4.2 Local Discourse Effects: Response Types and QAC . . . 8
  - 1.4.3 Higher Level Contextual Effects: Task and Role . . . 10
  - 1.4.4 Phonetic Detail: More Bang for Your Buck . . . 11
- 1.5 Chapter Synopsis . . . 12

2 Prosody and Dialogue Structure Notions 15
- 2.1 Introduction . . . 15
- 2.2 Prosody Basics . . . 16
  - 2.2.1 Basic Terminology . . . 16
  - 2.2.2 Representations on Intonation . . . 18
- 2.3 Dialogue Structure Basics . . . 21
  - 2.3.1 The Common Ground and Public Beliefs . . . 22
  - 2.3.2 Tasks and Questions . . . 24
  - 2.3.3 Dialogue Moves and Question Answer Congruence . . . 25
  - 2.3.4 Dialogue Act Schemas . . . 27

2.3.5 Where does prosody act? . . . 31
2.4 Previous Analyses of Final Rises . . . 31
2.4.1 Forward Dependence and Non-Finality . . . 31
2.4.2 Commitments and Contingency . . . 34
2.4.3 Tests, Modals, and the Common Ground . . . 36
2.4.4 Speaker Attitudes and Affect . . . 37
2.4.5 Iconicity and Universality . . . 39
2.5 The Way Forward . . . 41

3 Cue Words, Dialogue Acts and Prosody 44
3.1 Introduction . . . 44
3.2 Cue Word Responses . . . 46
3.2.1 Cue Words in the Switchboard Corpus . . . 46
3.2.2 Cue Words in SWBD-DAMSL . . . 47
3.2.3 Cue Words in the MDE Annotations . . . 51
3.2.4 Previous Work on Cue Words and Prosody . . . 53
3.3 Corpus Study: *Really*, Backchannels and Questions . . . 55
3.3.1 Goals and Hypotheses . . . 55
3.3.2 Data and Method . . . 57
3.3.3 Distributions in SWBD-MDE . . . 59
3.3.4 Prosodic Separability n SWBD-MDE . . . 63
3.3.5 Distribution and Separability in SWBD-DAMSL . . . 65
3.3.6 Summary . . . 68
3.4 Corpus Study: *Really*, DAs and Contextual Features . . . 69
3.4.1 Data and Method . . . 70
3.4.2 Results . . . 71
3.5 General Discussion . . . 74
3.5.1 Active/Passive Backchannel Questions . . . 74
3.5.2 About those rises... . . . 76
3.5.3 Dimensions of Prosodic meaning . . . 77
3.6 Conclusion . . . 78

4 The Interpretation of Cue Words and Rises 80
4.1 Introduction . . . 80
4.2 Perception Experiment: Cue Words and Speaker Attitude . . . 81
4.2.1 Data and Method . . . 81
4.2.2 Results . . . 82
4.2.3 Prosodic Features . . . 84
4.2.4 Subject Variation . . . 85
4.2.5 Discussion and Summary . . . 87

4.3 Perception Experiment: Terminal Rises and Uncertainty 88
4.3.1 Data 89
4.3.2 Method 90
4.3.3 Results 91
4.3.4 Discussion and Summary 96
4.4 Cue words, Gradability and Discourse Structures 97
4.4.1 Cue Words Across Corpora 98
4.4.2 Cue Words, Gradability and Evidence 101
4.5 Conclusion 112

5 Declarative Responses and Prosody 115
5.1 Introduction 115
5.2 Information Structure and Sentential Prosody 117
5.2.1 Pitch Accents and Metrical Structure 120
5.3 Information about propositions: Verum Focus 122
5.3.1 Verum Focus Responses 124
5.3.2 Previous analyses of VERUM 130
5.3.3 Arguments Against a Special Treatment of Verum 135
5.3.4 The IS View of Verum 138
5.4 Production Experiment: Declarative Responses 140
5.4.1 Goals and Expectations 140
5.4.2 Data and Method 143
5.4.3 Short Context Productions 144
5.4.4 Long Context Productions 154
5.4.5 Functional Principal Components Analysis 157
5.4.6 Separability of Prosodic features 161
5.4.7 Predictive Ability 164
5.4.8 Using the Classifiers to Explore Variation in the Data 167
5.5 General Discussion and Summary 170
5.6 Conclusion 173

6 Interpreting the Rise in a Fall-Rise 175
6.1 Introduction 175
6.2 Perception Experiment: Fall-Rise, Fall-Fall 177
6.2.1 Goals and Expectations 177
6.2.2 Data 177
6.2.3 Method 179
6.2.4 Results 180
6.2.5 Resynthesized Indirect Responses 182
6.2.6 Pair-list Responses 187

6.2.7 Discussion and Summary . . . 189
6.3 The Proper Analysis of the Rise in Fall-Rise Accent . . . 190
6.3.1 Fall-Rise and IS categories . . . 191
6.3.2 Intonation and Attitude: The Other Proposition . . . 200
6.3.3 Metalinguistic Uses . . . 210
6.3.4 Strategies, Rises and Coherence . . . 213
6.4 Conclusion . . . 215

7 Higher Level Dialogue Factors: Task, Role, Dialect 217
7.1 Introduction . . . 217
7.2 Background: Prosody and Higher Level Dialogue Features . . . 219
7.3 Data . . . 222
7.3.1 The IViE Corpus . . . 222
7.3.2 Additional Annotations and Features . . . 224
7.3.3 F0 Features . . . 226
7.4 Distributions of Moves in the Dialogues . . . 226
7.4.1 Type Frequencies . . . 228
7.4.2 Speaker Changes . . . 229
7.4.3 Summary: Move Characteristics of Task Oriented and Conversational Speech . . . 231
7.5 Boundary Intonation in the Dialogues . . . 231
7.5.1 Potential Connections between Move Structure and Intonation 231
7.5.2 Exploration of Intonational Features . . . 234
7.5.3 The Effects of Role, Move and Syntactic Type . . . 236
7.5.4 Turn-taking . . . 239
7.5.5 Discussion and Summary . . . 244
7.6 Dialectal Differences in Read Speech . . . 247
7.6.1 Shape differences . . . 248
7.6.2 Connection to ToBI style labels . . . 254
7.6.3 Beyond Labels . . . 257
7.6.4 Speaker Differences . . . 259
7.6.5 Discussion and Summary . . . 261
7.7 Conclusion . . . 265

8 Conclusion 267
8.1 Summary of Findings . . . 267
8.2 Where does intonation act? . . . 269
8.3 Future Directions . . . 271

A Production Experiment Materials **272**
A.1 Short Contexts 272
A.1.1 Direct Contradictions 272
A.1.2 Direct Agreements 273
A.1.3 Indirect Contradictions 273
A.1.4 Indirect Agreements 274
A.1.5 Direct Declarative Questions 274
A.2 Longer Contexts 275
A.2.1 Discovery/Affirm: Robbery 275
A.2.2 Contradictions: Gran's Birthday 276

Bibliography **278**

List of Tables

2.1 Levels of Grounding (cf. Benotti (2009)) 23
2.2 Most frequent SWBD-DAMSL labels 28
2.3 HCRC Map Task move percentages 29
2.4 Summary of rise analyses. 42

3.1 SWBD-MDE: *Really* DAs 57
3.2 SWBD-DAMSL: *Really*s DAs 58
3.3 Predicted DA from the curve feature logistic regression 63
3.4 SWBD-MDE cross-validation results 64
3.5 Cross-Validation results downsampling the data (151/151) 64
3.6 DAMSL: All data 67
3.7 DAMSL: Cross-validation results for downsampled data sets. 68
3.8 SWBD-MDE contextual features 70

4.1 Correlation coefficient (Kendall's τ) and p-values of the question/surprise ratings and prosodic features for *really* (top) and *right* . 83
4.2 *Means and standard deviations for question and surprise ratings, by subject* 85
4.3 *Pairwise Mann-Whitney U tests for question rating by subject with Bonferroni correction (** = $p < 0.01$, *** = $p < 0.001$).* 86
4.4 Cue word frequencies across corpora 98

5.1 First word of the highest level S for verum, non-verum utterances (SWBD-NXT/Treebank) 127
5.2 Identifiers for production target sentences 142
5.3 FDA: Proportion of Variance accounted for the first five principal components. 158
5.4 Last Word Aggregate Features 161
5.5 Legendre polynomial and FPCA (fda) based coefficients: Classification Results 163
5.6 Classification of long context productions 165

6.1 Steedman (2007): Tone meanings for English 197
6.2 Levels of Grounding 212

7.1 Accent Tones used in the IViE transcription Grabe (2004). Note: not all of these tones appear in the reported results. 224
7.2 Estimated standard deviation σ for group level parameters. 239

List of Figures

2.1 AM Finite State Automata 18
2.2 Nuclear H*L-H% or fall-rise? 19

3.1 SWBD-DAMSL: One word turns. 48
3.2 SWBD-DAMSL Dialogue acts for specific cue words. 49
3.3 SWBD-MDE one word SUs 52
3.4 SWBD-MDE SU types. 52
3.5 Legendre Polynomials 59
3.6 Means contours, by immediately previous cue word. 59
3.7 Parameter estimates for aggregate prosodic features and the immediate previous cue word (± two standard errors). 60
3.8 Parameter estimates for curve features. 61
3.9 MDE *Really*: F0 convexity v tilt 62
3.10 MDE *Really*: Height, Tilt, PCA 62
3.11 F0 height vs Intensity Convexity for (left) the whole data set and (right) a 'good' downsampled set. 65
3.12 *Really* in SWBD-DAMSL: F0 tilt (LC2) and convexity (LC3) 66
3.13 *Really* in SWBD-DAMSL: mean F0 and intensity contours 66
3.14 SWBD-DAMSL: *Really*, logistic parameter estimates 67
3.15 Two *really*s 68
3.16 SWBD-MDE context: Logistic regression parameter estimates 72
3.17 *Really* in SWB-MDE: Speaker Switches, evaluations 72
3.18 SWBD-MDE Context: Parameter estimates for group level predictors. 73

4.1 Average surprise versus question ratings. 84
4.2 Stimuli ratings ordered by mean average rating 86
4.3 Pitch range versus first syllable slope 87
4.4 Stylized pitch contours for *right*, with quantiles and contour mnemonics. 90
4.5 Mean scores for each cue word by question 92

4.6 Parameter estimation medians. The shaded range represents 2.5th-97.5th quantiles. Red: CREDIBILITY, blue: EXPECTEDNESS, green: EVIDENCE. 94
4.7 Cue word/contour interaction 95
4.8 An Optical Illusion 101
4.9 Evaluation of tall and short (Kennedy and McNally, 2005) 103
4.10 The quality thresholds of Davis et al. (2007) look like scales for gradable adjectives. 104
4.11 von Fintel and Gillies (2010) style kernel for epistemic modals. . . . 105
4.12 A generalized kernel for *really*. 107

5.1 Right branching metrical structure 121
5.2 Fall-rise in example (29): 'He DID steal the *money*...' 125
5.3 A classic verum update contour labelled as a question 128
5.4 Time Normalized Mean F0 (st) for short context productions 145
5.5 Verum focus 'did': Mean F0 (st). 145
5.6 Last word intensity and duration by response type 146
5.7 Shape Features: Comparing position of last contrast 147
5.8 Mean F0 Contours based on Legendre coefficients, last contrast . . . 147
5.9 Analysis of variance: Finite Population Standard Deviation (lw) . . . 148
5.10 F0: Response Type 149
5.11 Analysis of variance: Finite Population Standard Deviation (verum/did) 150
5.12 F0: agreement (did) 151
5.13 F0: speaker differences for utterance final 'meringue' (sentence E) . . 152
5.14 Intensity: speaker differences for utterance final 'meringue' (sentence E) 153
5.15 F0 z-scores by utterance: 'meringue' (sentence E) 154
5.16 Long vs Short Contexts: Mean F0 155
5.17 Contradiction Contour 155
5.18 Long and Short: ma2, sentence E 155
5.19 Direct (D) and indirect (I) verum responses in the long contexts . . . 156
5.20 FDA Mean and Harmonics: Time warped, whole utterance contour . 158
5.21 FDA Mean and Harmonics: Last words (fda.lw) 159
5.22 FDA projection: Last words (fda.lw) for short and long contexts (direct and indirect responses only) based on the short context production FDA. 160
5.23 First Three Principal Components 162
5.24 Visualizing the difference in prediction accuracy between seen and unseen data. 166
5.25 FDA Y predictions from the fda.lw classifier. Vertical lines indicate word boundaries, although this classifier only considers the data from the last word 'birthday'. 167

5.26 Projection of Y data for fda.lw 168
5.27 FDA pair-list predictions from the fda.lw classifier 169
5.28 Projection of pair-list data for fda.lw 170

6.1 Stylized resynthesis 178
6.2 Centered scores by question. 181
6.3 Different response types, PCA on scores. 182
6.4 Parameter Estimates. 183
6.5 Indirect Responses: mean scores and 95% bootstrapped confidence intervals. 184
6.6 Credibility 185
6.7 Non-finality 185
6.8 Sureness 186
6.9 Engagement 186
6.10 Pair-list (B) utterances 188

7.1 Proportions of moves for Cambridge speakers: $N_{conv} = 430$, $N_{map} = 1287$. 227
7.2 Sentence Types for Cambridge speakers 227
7.3 Affirmatives (Cambridge) 228
7.4 Proportion of speaker changes for both dialogue types 229
7.5 Speaker changes: Moves following Instructions by next speaker 229
7.6 Speaker changes: Moves following Informs grouped by next speaker . 230
7.7 Moves after affirmatives, by role. 230
7.8 Speaker changes after affirmatives 231
7.9 Inform and Instruct moves: Tilt and convexity. 232
7.10 Mean contours based on Legendre coefficients. 233
7.11 Tilt (LC2) and Convexity (LC3), by move type, with means. 233
7.12 Mean contours Instruct and Inform moves based on Legendre coefficients grouped by syntactic type. 234
7.13 Instructions overlaid with read declaratives 235
7.14 Inform utterances overlaid with read declaratives 236
7.15 Finite Population Standard Deviations: Medians 95% and 50% intervals 237
7.16 F0 height (LC1) parameter estimates. 238
7.17 F0 tilt (lcoeff.value.2) parameter estimates. 238
7.18 F0 convexity (lcoeff.value.3) parameter estimates. 239
7.19 Speaker change (ts) parameter estimates 240
7.20 With LC data as individual level predictors. 241
7.21 With LC data as individual level predictors and predictors on the move group. 242
7.22 Predicting Affirmatives after Inform and Instruct moves with LC data as individual level predictors and predictors on the move group. . . . 243

7.23 Nuclear Tune Proportions by region (cf. Grabe (2004)). 248
7.24 Nuclear Tune Proportions by move type from (cf. Grabe (2004)) . . . 248
7.25 Mean Contours by region (last word) 249
7.26 Mean Contours by sentence type (last word). 249
7.27 LC3 v LC2, by region (last word). 250
7.28 LC3 v LC2 with means, by sentence type (last word). 251
7.29 mean contours: whole utterance, by region 252
7.30 Last word and whole utterance means 252
7.31 IViE Tone Labelling: LC3 (convexity) v LC2 (tilt) and LC1 (height) v LC2 (tilt). 253
7.32 IViE Tone Labelling: LC3 v LC2 254
7.33 LC3 v LC2 by sentence type and tone: Cambridge and Leeds 255
7.34 H*L versus H*LH% in two Cambridge declaratives and one polar question 256
7.35 Declarative Statements (left), Belfast data (right) 257
7.36 Finite population standard deviation estimates 257
7.37 F0 height: LC1 model 259
7.38 F0 tilt: LC2 model 260
7.39 LC3 model 261
7.40 LC3 v LC2 by speaker (Newcastle). 262
7.41 Newcastle nm2 in dialogue. 262
7.42 LC3 v LC2 by speaker (Belfast). 262
7.43 bf3 in dialogue. 263

Chapter 1

Introduction

1.1 Prosody in Dialogue

This dissertation is about what prosody contributes to the interpretation of speech. That is, when, where, and why we get the systematic variations in prosody that we do and, most importantly, what they mean. More specifically, it's about the role and interpretation of terminal rises in discourse. The long term goal of this research project is to provide a robust and testable framework for integrating components of prosody, like rises, into models of dialogue. While the prosody-meaning mapping has been tackled from many different angles, the question of what exactly prosody does here still remains open. A major reason for this is that the types of data used to address this problem depend greatly on where the study sits on the spectrum of highly theoretical to purely application driven approaches. This has left a gap between accounts which deal with the high level mechanics of deriving prosodic meaning and those dealing with low level phonetic facts. This dissertation is an attempt to bridge that gap. Key questions that this work addresses are (i) what dimensions of meaning best reflect prosodic variation and (ii) how can we capture this variation quantitatively.

Theoretical analyses generally attempt to give an account of how intonational units, like terminal rises and pitch accents, relate to the semantics and pragmatics of their carrier utterance. This usually means giving intonational forms lexical entries. For example, the impressionistic link between rises and uncertainty has been analyzed in terms of possible world semantics (Nilsenova, 2006, Zaroukian, 2011). Similarly, specific pitch accent shapes have been identified with abstractly defined informational categories like topic or focus (Jackendoff, 1974, Steedman, 2000, Büring, 2003). These approaches have an easier time integrating with existing formal frameworks for di-

alogue. However, such theoretical accounts are generally qualitative in nature and often based purely on author intuitions. This has led to mappings between prosodic form and meaning which are too tight and so do not scale up well when we look at more data.

More data-driven approaches have also been attempted, particularly with respect to the question of how prosody relates to speaker attitude and emotion (Litman et al., 2009, Pon-Barry, 2008, Amir et al., 2010) and dialogue acts (Shriberg et al., 1998, Gravano et al., 2007). These studies take a more quantitative approach to the analysis of prosody and are usually based on features derived from the speech signal, e.g. aggregate statistics about fundamental frequency, energy, and duration, or otherwise human labelled approximations of prosodic forms, e.g. ToBI annotations (Silverman et al., 1992). As such, the analyses mostly describe the distribution of features with respect to attitudinal or discourse categories. While this approach is informative, it avoids the crucial question: what are those features doing there in the first place. This is something we need to know if we want to tackle, for example, the problem of synthesizing expressive, conversational speech.

To really understand what prosody contributes to spoken discourse we need to look at how prosody varies in quantitative way, but we also need to understand the high level mechanics of how prosody, semantics and pragmatics interact. Moreover, we'd like to avoid manual annotation of intonational form and go straight to the signal to see what is really going on at the prosodic ground level. To do this, we need to know what variation in prosody is important for the interpretation task. This in turns requires us to identify the infrastructure that we actually need to interpret discourse and prosody. That is, to understand the contribution of prosody, we have to understand the things that it acts on.

I argue that we can get a better grip on the problem if we look at discourse structure as fundamentally task driven. From this perspective, we see that conventionalized aspects of prosody like terminal rises are used to situate utterances within a discourse structure. Moreover, they are used to influence expectations about what sort of structure should come next. The link to attitudes like uncertainty then do not come directly from the prosodic form, but instead, derive from a number of contextual factors. The most important of these is how the utterance fits with what is projected by the current discourse structure (cf. Farkas and Bruce (2010)). I attempt to show that the interpretation of these prosodic elements is mediated by such discourse expectations in a predictable way. In general, prosody signals the attitude of the speaker *towards the discourse* conditional on those expectations. More specifically, terminal rises signal whether the dialogue has come to a viable stopping point when considering the task at hand.

My case is built on a detailed examination of cue word and declarative responses

in dialogue. We will see that these types of responses can span a range of meanings from exclamative to questioning to confirmation. To get to these interpretations we need to consider not only the underlying semantics and prosodic form of the response itself but also the effect of global and local discourse context. To investigate the effect of higher level features we look at further data which varies in terms of dialect, task, and speaker role. These studies show the utility of using data analysis techniques that give us a more direct view what prosody is doing more directly and allow us to see more of the phonetic detail on larger sets of data. The rest of this chapter gives an overview of the data investigated and the central arguments made in this dissertation. But before we get to that, I layout the basic research questions and background about previous approaches.

1.2 The Research Problems and Background

The big question that this research project attempts to make progress on is:

(1) **What does prosody contribute to discourse interpretation?**

This dissertation attacks this problem by examining a widely used and highly identifiable prosodic element: the terminal pitch rise.

(2) **What do terminal rises contribute to discourse interpretation?**

The basic data point about rises in English is that statements fall while polar questions rise. As such, a lot of work on rises has been driven by a desire to explain the intuitive association between rises, questions, and uncertainty. From there, terminal rises have run the gamut of analyses with respect to what and how intonation contributes to meaning. On one end, we have Gussenhoven's (2002) biological codes, extending Ohala (1984), which makes a pre-linguistic connection between high pitch and a submissive stance. On the other more formal end, Nilsenova (2006) models the uncertainty associated with rises by giving it a modal meaning within a dynamic semantics framework. Similarly, Zaroukian (2011) gives a possible worlds semantics to rises, analyzing them as expressing contingency (cf. Pierrehumbert and Hirschberg (1990) and Gunlogson (2008)). Reese (2007) assumes that rises express uncertainty, but places this at the level of cognitive modelling in the framework of Segmented Discourse Representation Theory (Asher and Lascarides, 2003).

We can detect some broad themes amongst these accounts, e.g. the connection between rises and some sort of discourse contingency. However, these accounts differ

greatly in what they assume the domain and the range would be for a prosody-meaning mapping. So, a core question that needs to be addressed is:

(3) **What dimensions of meaning do terminal rises (and prosody more generally) act on?**

Surveying the territory, we find two main approaches to addressing this question which I will broadly group as 'lexical' and 'structure indicating' approaches. The former focuses on mapping intonational shapes to specific meanings, so that they contribute information, additional to the proffered content of the utterance, in a similar way to a sentential adverb or conventional implicature (Ward and Hirschberg, 1985, Nilsenova, 2006, Constant, 2007). On the other hand, the structural camp take prosodic forms as a signal of how to put linguistic units together at various levels, e.g. phonological, syntactic, information structure, discourse (Selkirk, 1995, Pierrehumbert and Hirschberg, 1990, Steedman, 2000, Calhoun, 2010, a.o.). The disparity between these accounts is reflected in the various different response variables, i.e. the categories and scales, that prosody has been linked with. To figure out what is actually going on, we need to figure out which response variables most directly reflect the variation in prosody. For example, do we really want to try to make a direct connection between prosodic features and dialogue acts or attitudes? Ideally, we want to be able to express this in terms of existing tools of semantics and pragmatics.[1]

However, part of the difficulty in resolving these issues is that these sorts of theories are usually built upon a foundation of intuition. This usually means that the analysis is based on isolated utterances or a limited inventory of contexts. For example, Büring (2003), following Jackendoff (1974), argues that fall-rise accents mark contrastive topics based on its appearance in answers to wh-questions. These studies are qualitative in nature so its hard to see how much coverage they actually provide. More quantitative studies on larger data sets have also been able to make link between rises and categories in certain types of dialogue situations. For example, Pon-Barry (2008) and Litman et al. (2009) have found rising pitch to be an indicator of uncertainty. However, these studies do not really tell us why we get rises when we do or whether the rise is really meaningful. For example, Benus et al. (2007) find that rises are common with affirmative backchannels which are, in general, understood to not express uncertainty. It seems that to get to the right tools for the job, we'll have to more closely at the components of a discourse.

(4) **What in the discourse context affects the interpretation of prosody? How do we model it?**

[1] In this sense, this sort of question falls into the same ballpark as other theoretical investigations of 'extra' dimensions of meaning, e.g. conventional implicature in the style of Potts (2005).

As there are many accounts of what terminal rises contribute, there are many accounts of how to model single and multi-party discourses. If we are going to unravel the effect of discourse context, we need to know what structures to model this with. Do we want to build a theory around, for example, rhetorical relations (Asher and Lascarides, 2003, Reese, 2007), or a question-answer structure (Roberts, 1996, Ginzburg, 2012, Farkas and Bruce, 2010). In reality, discourses have both. However, the key thing for us is which one is better for explaining what is happening with the prosody. We also want to know what the relationship is between such representations of discourse and the mechanisms and structures they are built around. That is, we want to articulate the relationship between prosody and notions like common ground, public beliefs, and the question under discussion.

In parallel to having the high level machinery in place, we need ways to be able to relate impressionistic assessments of prosody with what is actually going on in the speech signal. The flipside of finding appropriate response variables is that we need to evaluate how well suited our current techniques for dealing with the prosodic manifestation are for this task. That is:

(5) **How can we analyze differences in prosody quantitatively?**

As mentioned previously, many empirical studies of prosody are based on aggregate statistics about pitch, intensity, spectral features, etc. This contrasts with qualitative studies which talk about prosody in contour terms. While a characterization of the data in terms of those aggregates is often instructive, we would like to be able to bring these approaches together. Several larger scale investigations have used the ToBI framework in order to move towards this (e.g. Rosenberg, 2009, Gravano, 2009, Liscombe, 2007). This annotation scheme has its origins in Pierrehumbert's (1980) Autosegmental-Metrical treatment of English intonation and so the assumption is that this sort of tone labelling scheme reflects the actual phonology of English. However, experimental studies like Dilley (2010) suggest that the current pitch accent inventory does not actually represent the phonological and phonetic facts. Moreover, tone labelling using the ToBI annotation scheme has notoriously low interannotator agreement (Syrdal and McGory, 2000). So, reliability and reproducibility are at issue when using this sort of data. As such, tone labels look like a not-quite-consistent screen through which to view the prosodic action. In fact, labelling of tones has been abandoned in favour of a prominent/unprominent type labelling in recent prosodic annotation effort, e.g. Ostendorf et al. (2001).

However, there are still two parts to tone labelling that are not usually captured in empirical studies otherwise. The first is the identification of the time domain of the prosodic event of interest. The second is a description of the shape of the pitch

contour through that point of interest. Capturing this requires something more than pitch means and slopes over words or syllables. We would like to look at this without manual annotation. In this dissertation, I will try to show that this is possible if we have enough discourse information at our disposal. Moreover, I argue that it gives us a clearer understanding of what's going on. However, to do this effectively, we need to know what sort of prosodic variation actually makes a difference for interpretation. Considering the case of terminal rises, we get to the next question:

(6) **Do differently shaped rises mean different things?**

In particular, we would like to know whether ending high, i.e. having a high boundary tone, is independently meaningful of what comes before it in the prosodic context, as suggested in Pierrehumbert and Hirschberg (1990). For sentential utterances, this will require us to look at the interaction of boundary tone and pitch accent.

To find out what is important for interpretation we also want to take into account variations due to higher level contextual features. For example, speaker role, the nature of the task, and dialect. There appear to be differences in prosody between types of discourses, e.g. task oriented versus conversational dialogue versus monologue. Similarly, some studies have suggested that rises associate with particular speaker roles. For example, the analyses in Merin and Bartels (1997) and Steedman (2000) suggest that rises place the speaker in something of a submissive role. However, Cheng and Warren (2005) find a higher frequency of rises by speakers in a more dominant role in specific situations. So, the link between rises and role in this way is not clear.

More generally, we want to know if the meaning of rises stays the same in these different situational settings. If rises do have a consistent meaning across situations, we would want to be able to derive other the shades of meaning from the discourse structure and utterance properties in a similarly consistent fashion. We would also like to know how abstract the relationship between terminal rises and its associated meanings really is. For example, if we were to attach meaning to a high boundary tone for Standard American English, for example, would it be just as easy to associate that meaning with a low tone in a different system, or is prosody really iconic at heart. We might expect to see consequences of this in dialects which have final rises as the default boundary movement on statements, e.g. Belfast English (Grabe and Post, 2002).

1.3 The Data

The bulk of the analysis present in this dissertation is built from investigations of the prosody of cue words and declaratives responses in dialogue. The motivation for focusing on responses is that they provide better insight into dialogue structure than the types of utterances usually examined in theoretical accounts of meaning. Linguists tend to focus on utterances that make sense 'out of the blue', e.g. asserting something, asking a question, requesting an action. These canonically correspond to basic sentence types: declarative, interrogative, imperative. This sort of approach basically restricts one's point of view to forward looking uses of these constructions (Core and Allen, 1997). However, I would argue that taking the backwards looking view coaxes the infrastructure into the foreground, and it is this sort of structure that we are really interested in. When you take this view you have to deal with more structural considerations, e.g. Question answer congruence and, more generally, information structure. You also get a better line of sight into the expression of speaker attitude. For example, expression of surprise seems to signal a gap between the current state of the discourse and speaker expectations.

The first part of this dissertation examines cue word responses with data drawn primarily from the Switchboard Corpus of conversational telephone speech (Godfrey et al., 1992). Cue words, like *okay*, *yeah*, *right* and *really*, are discourse markers which appear frequently in spoken dialogue. The fact that they have specific segmental forms and lexical semantics makes it easier to study the effects of prosody on their interpretation. Indeed, there their interpretation seems to depend on prosodic factors (Benus et al., 2007, Gravano et al., 2007). However, what exactly prosody contributes and how this relates to theories of prosody and meaning still remains to be fleshed out.

Cue words are a good test subject for investigating prosody, but they also play a highly conventionalized role in dialogue. We would like to see if the conclusions we draw about their prosody extend to sentential responses. Hence, the second part of this dissertation focuses on declarative responses. The data comes from a set of scripted dialogues between pairs of speakers. To control for information structure of the sentential responses, we will look specifically at declaratives with a set IS structure. To keep things close to the discourse/propositional level I present production and perception data on broad focus and verum focus utterances. The former present the whole proposition as IS Focus, while I argue that the latter present narrow IS focus on polarity and are 'about' the proffered proposition. By controlling this, we can better examine the interaction between discourse structure, information structure, pitch accent placement and overall rise shape. To examine the effect of discourse structure, we will look at their prosody given varying amounts of dialogue context.

Finally, I examine the prosody of data from the corpus of Intonational Variation in English (IViE, Grabe and Post (2002)). This data set includes read speech, conversational speech and map task dialogues speakers recorded from various regions in the United Kingdom and Ireland. The goal here was to look at the effect of high level contextual features like speaker role, task and dialect on the interpretation of rises. In particular, we want to know how robust the conclusions about the nature of prosody are across different dialogue types. Portions of the IViE data have ToBI style prosodic annotation, so we can also compare the results from that sort of analysis to our analysis of the phonetic detail.

1.4 Key Claims

1.4.1 Prosody, Discourse Attitude and Structure

In this dissertation, I argue that conventionalized aspects of prosody, in particular intonation, act at the level of discourse structure. I claim that terminal rises signal that a discourse has not come to a viable stopping point. More generally, I argue that terminal rises and pitch accents should be treated as structure indicating rather than providing extra lexical style content.

In this sense, rises signal an attitude towards the state of the discourse rather than a attitude towards the proposition alone. So, the extent to which we can associate perception of an attitude with an intonational shape really depends on the how the utterance fits into the discourse. Only when the discourse context is willing do rises give the feeling that the speaker is *uncertain* about the Gricean Quality of their utterance. Since cue words have quite a conventional role in dialogue, we find that the main constraint in this respect is the underlying meaning of the cue word itself. Since there is infinitely more variability in meaning for sentential utterances, we have to pay more attention to the surrounding to the discourse structure to figure out what the prosody is doing there.

1.4.2 Local Discourse Effects: Response Types and QAC

At the local level, we see that prosody affects discourse interpretation of sub-propositional units. Whether a constituent is identified as a particular type of Information Structure (IS) unit depends mostly on what is projected by the immediately proceeding discourse. That is, the crucial guide for how local context affects the interpretation of intonational components are the expectations arising from Question Answer Congruence (QAC) (Büring, 2003). So, to understand what prosody is doing

we are better off looking at dialogue in terms of the current task/question under discussion, i.e. a stack driven model (Roberts, 1996, Ginzburg, 2012, Farkas and Bruce, 2010) rather than a graph of rhetorical relations (Asher and Lascarides, 2003, Reese, 2007). To model how this task driven structure gets glued together (which is what intonation is about, really), we also need to track individual participants' public commitments and responsibilities, as well as their shared commitments (Portner, 2007). However, for sentential utterances, it's the QAC that is crucial for determining what discourse interpretation a response gets.[2]

Prosody and Response Types

The study of declarative responses shows distinct intonational forms based on whether the utterance directly or indirectly addresses the question under discussion, i.e. has the expected QAC. We also find a distinction between evaluations (e.g. contradictions, affirmations) and checks. Even if we only consider declaratives, contextual constraints result in different prosodic forms. With this data in view, I claim that these sorts of *response types* are the kind of category which reflect what prosody is doing. While there is some reflection of the classical dialogue or speech act in this, i.e. adding new information versus requesting a check, this sort of category is not represented in prevalent in dialogue act annotation schemata. The reason that response types do a better job for us is that they are more accurate reflections of the current and projected discourse structure than your usual notion of dialogue act.

The differences in the production data are most succinctly categorized by the shape of the nuclear accent tail: direct responses are characterized by final falls, indirect responses come with a fall-rise while the checks have a rise proper (i.e. a concave shape accent). The nature of the rises on indirect responses and direct checks (i.e. declarative questions) highlights how prosody varies with response type. Both types are amenable to terminal rises, but vary in the *prosodic context* provided for those rises. From here it seems that a rise proper often implicates dependency with the hearer (Steedman, 2000, Gunlogson, 2008). The fall-rise contours in indirect responses generally don't generate the same sort of hearer dependency: the speaker does not seem to defer to the hearer on the issue of adding the proffered content into the common ground. Nevertheless, terminal rises are congruent with indirect responses because they appear to point to an open question under discussion.

[2]Note: this is basically the part of Farkas and Bruce (2010), i.e. projected common grounds, that differs from other QUD centric accounts like Ginzburg (2012) and Roberts (1996)

Appeasing the QAC: Strategies and Rises

I argue that the indirect responses (in the production experiment) have the broad intonational pattern they do because they invoke a strategy for addressing the current question under discussion via subquestions. The fall-rise accent marks an element of the IS ground as contrastive, which gives the hearer a template of what sort of questions the strategy contains. The crucial point is that while a fall-rise accent is good for the job, there doesn't have to be a rise for this to work: there just needs to be contrast. Invoking a strategy maintains question answer congruence by basically deferring evaluation with more questions. So, these indirect responses match the general meaning of terminal rises: this is not the time to stop the dialogue. This view is supported by the perception data. Indirect responses were generally evaluated as signalling that more information is required, regardless of the fall/rise characteristic. As with cue word *really* and interrogatives, the signal coming from a terminal rise is being strongly signalled from elsewhere in the utterance and so is somewhat redundant. Still, the rise provides information: it's a bigger gesture and so it makes the speaker sound more involved/engaged – the speaker actually wants the issue to be closed and closed with consensus. This leads to the *politeness* that has been associated with rises.

Overall, the intonation of an utterance that maintains discourse coherence, as guided by question answer congruence. If the content has the form of a (relevant) direct answer to a question but the prosody has a rising accent, the hearer will probably assume that the speaker has left the question open because they have some uncertainty about whether the answer is true. If such accommodations can't be made or the interlocutor requires greater precision than has been given,[3] the dialogue may then go into a repair mode where clarification of the common ground is becomes the top priority task.[4]

1.4.3 Higher Level Contextual Effects: Task and Role

When we look at the relationship between rises and higher level context, we see that utterance final rises become more prevalent as speakers become more interactive and the need to co-operate increases. We see this when comparing monologue and dialogue (you get more rises in the latter). This makes sense given that monologues are by nature not as co-operative or interactive. Conversational dialogue involves the construction of common ground and so requires more signals to co-ordinate the

[3]That is, a higher Quality Threshold (Davis et al., 2007).

[4]In the case of the object wh-question, with a subject focus response, but where the answer is still the answer, the hearer will probably ask 'why did you say it like that?'.

interaction. The need to synchronize the common ground is even more pressing in instructional/collaborative tasks, like the map task. In this case we expect to find more confirmation eliciting rises, since moving forward in the task depends on the participants being at the same place. This is indeed what we see in the IViE data comparing read, conversational and task oriented speech. By signalling that the others should confirm what they are saying, rises in the forward checking mode reinforce that a particular speaker is directing the dialogue. So, while terminal rises do seem to signal a desire to co-operate and synchronize, they don't necessarily express uncertainty or submissiveness (cf. Merin and Bartels (1997)).

1.4.4 Phonetic Detail: More Bang for Your Buck

Finally, looking at the variation in the data, it becomes apparent that the phonetic detail matters. This is highlighted in the IViE dialectal data, where we see that both declarative questions and statements in Belfast English end high, but the former have a higher rise. So, to distinguish these types we either have to make a distinction between these different rises, or consider tone plus a scaling factor. Either way, we have to take measurements from the signal so its not clear what we gain from a tone labelling intervention that we couldn't get from looking at contour shapes directly modulo normalization. In fact, our contour analysis on the IViE data basically include and extend reported results based on tone labelling (Grabe et al., 2005).

Similarly, the distribution of pitch contours of cue words from the Switchboard Corpus, like *really*, seem to show a continuous deformation from rising to flat to falling. It seems that trying to make categorical pitch distinctions here misses something important with respect to interpretation in dialogue. For example, perception studies indicate that how questioning *really* is perceived to be correlates with the size of the pitch gesture. We can view this as a combination of high speaker engagement and the underlying check semantics of this cue word. So, the size the gesture has implications for what happens next in the dialogue, even if the rise characteristic doesn't in this case.

More generally, if we want to to analyze prosody quantitatively, rather than looking for prosodic events in and of themselves, it seems we should be identifying utterance components that are relevant for discourse structure and then analyzing the prosody over those units. This involves identifying units in the expected IS structure projected by the QAC and, similarly, potential points of contrast. Annotation of IS structure is still, of course, extremely labour intensive and automation of this is in its infancy. However, I believe this sort of human intervention will be more fruitful in the long run for both the study of prosody and discourse structure in general.

1.5 Chapter Synopsis

- Chapter 2 provides lays out the basic notions about prosody and dialogue structure I will be using in the rest of the dissertation. I review previous accounts of rise meaning, highlighting the different parts of discourse structure these analyses assume rises act on. Trying to mesh the results of previous studies on prosodic meaning suggests that rises should be analyzed in terms of the current questions under discussion, in a task driven model of dialogue along the lines of Roberts (1996) and Farkas and Bruce (2010).
- Chapter 3 examines the distribution of high frequency cue words in corpora of spontaneous speech. In particular, we will look at the relationship between the prosody of *really* and its associated dialogue acts though a corpus study. *Really* is sometimes interpreted as a question (i.e. something to respond to) and sometimes as a backchannel (i.e. something that can be ignored) which makes it a good probe for the question of how prosody affects interpretation in a dialogue. I attempt to quantify the prosody of these utterances through various aggregate statistics as well as functional decomposition methods. We will see that while some instances provide the canonical rising or falling shapes one might associate with tone labels, the distribution of contours has a fairly continuous spread in the function space. Classification experiments indicate that the variation in the prosody of *really* doesn't line up with the meaning suggested by this sort of dialogue act categorization. In particular, we don't get find that rises associate with questionhood any more than falls.
- Chapter 4 presents perception experiments with the goal of finding a better characterization of how the prosodic variation on cue words affects their interpretation. The results suggest that we can get a clearer indication of what the prosody is doing if we look at perceived speaker attitude. In particular, for *really*, expanded pitch range on both rises and falls intensifies the perception of speaker surprise and questioning. However, looking at a range of cue words, we find that this attitude is highly dependent on the underlying semantics of the cue word. So, rather than adding independent attitudinal meaning, a better characterization for terminal rises is that they signal that more needs to be said about the current question under discussion. From a methodological point of view, the results of these experiments suggest that we can use attitudinal scales to investigate the prosody-meaning mapping, but only if we have a grasp of the underlying semantics and pragmatics of the utterance.
- Chapter 5 presents a production study of declarative responses in varying dialogue contexts. The main goal of this was to see if the generalizations about terminal rises made with respect to cue words extend to sentential utterances.

To account for information structural considerations we look specifically at utterances with broad and verum (polarity) focus. Aside from the matter of terminal rise interpretation, additional analysis of these propositional level focus types contributes to our understanding of the relationship between pitch accents and information structure and their relationship to discourse structure. With this is in place, I find that certain response types do elicit specific intonational forms. Unlike the cue word/dialogue act case, machine learning classification experiments show that we can separate out the different response types based on aggregate statistics like mean F0. However, we will see that we get better performance predicting response types of new data when we use more shape oriented features, e.g. coefficients from Legendre polynomial decomposition or functional principal components analysis.

The production experiment data highlights the fact that indirect responses elicited fall-rise type tunes. While IS clearly plays a role in the intonational shape of an indirect response, fall-rise accents can appear in the IS focus and the ground of an utterance. This casts doubt on accounts which attempt map this particular accent shape to a specific IS type (cf. Büring (2003)). Moreover, while the fall-rise was prevalent in the production data, it didn't always appear. This suggests that the rise is not as important for matching response type as the marking of an informational unit as contrastive. This makes sense if we take this extra contrast as evoking a *strategy* (Roberts, 1996, Büring, 2003). This signals that some related subquestions should be used to resolve the current question under discussion. This in turn implies that the discourse has not reached a good stopping point, independent of whether a rise is there or not.

- In this vein, Chapter 6 investigates whether terminal rises contribute additional (contentful) meaning to indirect responses. To do this, we present a perception experiment of declarative responses varying pitch peak height and terminal rise/fall characteristics. The results of the perception experiment indicate that the presence of a rise does not increase the perception of uncertainty or discourse openness on top of what is already given by the response type. Adding a rise does seem to enhance the perception of speaker engagement – the rise makes the gesture bigger, more articulated. This data helps us sort out the myriad analyses of the meaning of the fall-rise contour extant in the literature, falling down on the side of the strategy based approach of Büring (2003), but rejecting its tight mapping between *that specific accent shape* and the category of contrastive topic.
- While local discourse structure clearly has a big effect on how prosody is interpreted, we also need to look at the effect of higher level features. To do this, Chapter 7 presents series of studies based on the IViE corpus of English dialects.

These studies examine the difference in the manifestation and distribution of terminal rises given factors such dialect, task, and role. Examination of task-oriented and conversational speech showed more rises on proffering moves and affirmative responses in the task oriented speech. In general, we can see this as a reflection of the fact that participants needed maintain a greater level of co-ordination in the map task.

Looking at dialectal differences we see that while Belfast English has default rising statement, it is not a mirror image of 'falling' Southern British English – declarative questions still rise, they just rise higher. Conversely, declarative questions in Southern British English can be falling, but have bigger pitch accents than their statement counterparts. Analysis of the data makes it clear that these differences can only really be seen by looking at the phonetic detail, particularly contour shape, directly. This difference is styles highlights the interactive function of rises, even in 'rising' dialects. I argue that, taking interactivity and the nature of the task into account, we can explain some of the contradictory accounts of the effect of rises. For example, by analyzing rises as primarily directing discourse structure, we can see why they would have been associated with both dominant and submissive roles in *specific* contexts (cf. Cheng and Warren (2005)).

- Chapter 7 presents general conclusions of the study and presents issues for further consideration.

Chapter 2

Prosody and Dialogue Structure Notions

2.1 Introduction

This chapter is about establishing the basic prosodic and discourse structural notions we will be working with in the rest of the dissertation. The problem of prosodic meaning has been attacked from many angles. With these different angles come a plethora of representational differences. Looking at these differences highlights methodological differences through which prosody and meaning have been studied. From the theoretical perspective, the prosody/meaning connection has generally been investigated in categorical terms, e.g. high versus low boundary tones, fall versus fall-rise accents. Unsurprisingly, the output of these analyses tend to be categorical in nature as well. Most saliently, various semantic/pragmatic studies have attempted to provide lexical entries for items from the tonal inventory of Autosegmental Metrical (AM) analyses of English intonational phonology (Pierrehumbert, 1980, Pierrehumbert and Beckman, 1988) in terms of how they act on the meaning of their carrier utterance.

More empirical work on the prosody and meaning has come from the analysis of speech corpora. This has often been done with the goal of improving voice applications. As such, such studies tend to deal directly with acoustic features. However, several corpus studies have been based on human annotation on intonation, the ToBI (Tone and Break) Indices labelling scheme (Silverman et al., 1992), deriving from the AM framework being the the dominant annotation method in that regard. In either case, the output is usually a description of the distributions of those features conditional on discourse categories like dialogue acts or attitudes like uncertainty. The generalization then comes in the form of a stylized fact about the data rather

than a licensing condition or semantic operator. As such, it is not always easy to see the implications of findings from instrumental studies for theoretical accounts of prosodic meaning. In order to see how the empirical work bears on the more theoretical approaches, we need to look at how prosodic representations relate to that of dialogue. That is, we need to get an idea of the type of information and structures that dialogue participants need to keep track of and how these structures change during a dialogue. From there we can get a better idea of where and how prosody acts in dialogue interpretation.

This chapter is structured as follows. Section 2.2 reviews basic notions about prosody and its representation. Section 2.3 explains the model of discourse structure we will use to talk about the role of prosody in dialogue. Section 2.4 discusses previous analyses of rise meaning in these terms. Section 2.5 recaps the issues that we will need to address in the rest of the dissertation to get a grip on what rises contribute to dialogue interpretation.

2.2 Prosody Basics

2.2.1 Basic Terminology

In this dissertation, I will use the term *prosody* to refer to the suprasegmental phonetic features of an utterance. These are usually characterized in terms of *pitch*, *loudness* and *timing*. In terms of acoustics, pitch is characterized as a psychophysical correlate of the *fundamental frequency* (F0 in Hz) of the sound wave. The subjective notion of loudness is usually expressed in terms of intensity, measured in decibels (dB), derived from the wave pressure amplitude. Timing is usually analyzed in terms of the duration of various speech units, e.g. segments (consonants, vowels). For prosody research the basic timing unit is usually taken to be the syllable. Syllables are grouped into higher level prosodic units such as *prosodic words* and *phonological phrases* (Shattuck-Hufnagel and Turk, 1996). Following Ladd (2008) and Calhoun (2007), I assume that phonological phrase structure is defined recursively. Voice quality has been argued for as the 'fourth prosodic dimension' (Campbell and Mokhtari, 2003). This refers to manner of phonation, e.g. creaky, breathy, or modal, which is often characterized in acoustic terms through the difference in the first and second harmonics in the Fourier spectrum.

These suprasegmental features have been often used to characterize prosodic *prominence*. Prominence is a relative term: a prominent syllable is more salient than those around it. This term is intimately bound up with notions of *stress* and *accent*, a distinction originally made in by Bolinger (1958) and re-introduced in Pierrehumbert

(1980). A *stressed* syllable is one that will be made prominent, i.e. *accented* if the word is emphasized. The term *pitch accent* refers to the idea that prominence is realized as some sort of pitch change on a syllable. However, Bolinger notes that there several other phonetic cues to accent, e.g. vowel quality, duration, and loudness. Still he suggests that 'pitch usually carries the day against length and loudness' (Bolinger, 1986, pg. 22). However, several other studies have highlighted the utility of non-pitch correlates, e.g. spectral tilt, in studying prominence (Sluijter and Van Heuven, 1996, Campbell and Beckman, 1997, Kochanski et al., 2005).

I reserve the term *intonation* to refer to prosodic features describing post-lexical use of pitch. This is distinguished from lexical *tone* in languages such as Mandarin Chinese, where pitch patterns distinguish different words. The study of intonation in terms of abstract representations is the domain of *intonational phonology*. The goal of such investigations is to characterize the sounds of an utterance 'in terms of a small number of categorically distinct entities' (Ladd, 2008, pg. 10). In general, utterance intonation has been characterized by two types of intonational event: pitch accents and phrase terminal pitch movements or *boundary tones*. The main work of intonational phonology is about the characterization of such events in terms of their shape and their placement.

The placement problem for pitch accents is intimately intertwined with issues of *metrical structure* (Liberman and Prince, 1977, Selkirk, 1986). The metrical structure of a phrase of a phrase determines the relative strength, hence prominence, of syllables. This is represented through a tree structure over syllables where sister nodes display a binary weak-strong alternation This basic weak-strong alternation is the basis for the perception rhythm in speech. The prominence of a syllable is then *relative* to the expectations projected by the metrical structure of an utterance. The *nuclear prominence* (i.e. nuclear accent) falls on the word/syllable that is dominated by only strong nodes, i.e. the most *structurally prominent* position. The assumption is that English has a default right branching structure, so the default nuclear prominence/accent position is at the utterance end. However, the actual mapping between strings of words and metrical structure is dependent on multiple linguistic factors, e.g. phonological, and syntactic, and information structural considerations (Calhoun, 2007).[1]

Given that we have identified an intonational event, we come to the problem of whether the shape of the pitch contour at that point means anything. That is, whether having a concave ('peaked') shape on a prominent syllable means something different from having convex ('scooped') one. The object of our current study, terminal rises, clearly falls under this more general problem. Several works have posited relationships between intonational forms and attitudes or structural categories, with varying

[1] We will look at the role of information structure in more detail in Chapter 5.

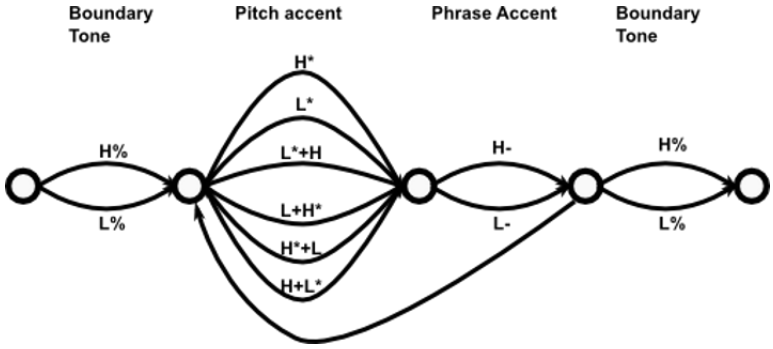


Figure 2.1: AM Finite State Automata

levels of formality (O'Connor and Arnold, 1961, Pierrehumbert and Hirschberg, 1990, Gussenhoven, 1984, Steedman, 2000, a.o.). In order to be able to determine whether such prosody-meaning maps stand up to scrutiny, we need to know a little about the domain such maps take. That is, we need to take a look at how intonational form is represented.

A major difference in how intonational events are represented is between whether pitch accents and boundary tones are treated as separate tonal targets, or whether the contour is treated as single unit. The former is indicative of from the *autosegmental metrical* (AM) approach of Pierrehumbert (1980), while the the latter characterizes descriptions from the *British tradition* (cf. O'Connor and Arnold (1961)). These provide categorical descriptions of contours, which are perceptually derived. Another approach is to directly model intonation (or prosody more generally) from acoustic features. We review these different approaches in the next section.

2.2.2 Representations on Intonation

The current dominant approach for describing intonation in English is based on the AM framework of Pierrehumbert (1980) and Pierrehumbert and Beckman (1988), where contours are described as sequences of high (H) and low (L) tonal targets. These tones are associated with three phonological elements: pitch accents (X*), phrasal accents (X-) and boundary tones (X%). Phrasal accents and boundary tones can only take single H and L values. Pitch accents, however, can be bitonal. So, besides the central tone of a pitch accent (H*, L*), a pitch accent can have leading and trailing H and L tones. For example, a L+H* accent describes a concave accent with a leading low target, resulting in a more distinct F0 rise before the peak. Admissible tone sequences are described by a finite state automated as shown in Figure 2.1. Phrasal accents align with ends of intermediate phrases and describe the pitch contour

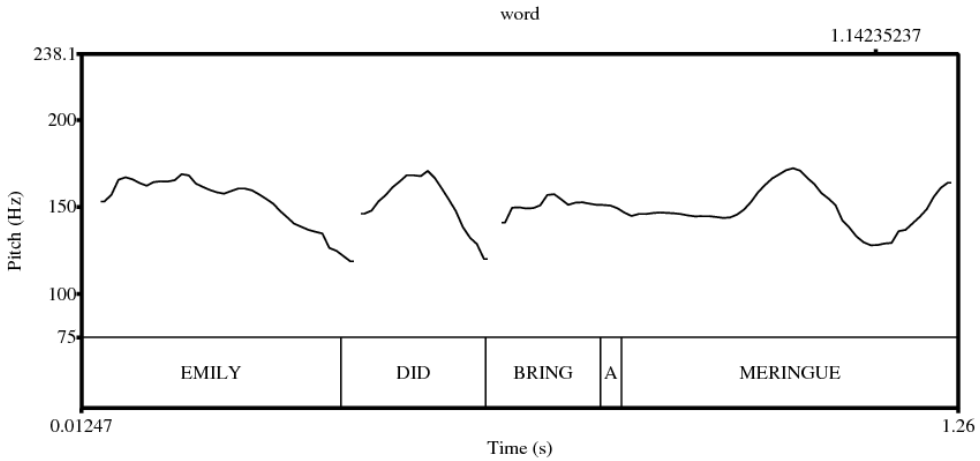


Figure 2.2: Nuclear H*L-H% or fall-rise on *meringue*?

between the accent and the intonational phrase boundary (where boundary tones align). The nuclear accent is defined as the last accent of an intermediate phrase.

Instead of using level tones, contours are described in British tradition in terms of dynamic characteristics like falls and rises, i.e. configurations (O'Connor and Arnold, 1961, Crystal, 1969, Halliday, 1970, Brazil, 1997, Cruttenden, 2007). These falls and rises implicitly indicate points of inflection: we take a fall to mean a pitch accent that falls from a peak. Descriptions generally focus on the contour from the nuclear prominence to the end of the intonational phrase, and so pitch accents, phrase accents and boundary tones are described together. Figure 2.2 shows what would be described as a fall-rise in the British tradition, but as a H*L-H% combination in the AM framework. We can view the AM mode of description as an *on-ramp* sort of analysis: it describes the direction of the contour leading up to the point of inflection. The British tradition, on the other hand, can be thought as an *off-ramp* characterization, describing the shape of the contour after the point of inflection.

For purely descriptive purposes, there is not a lot of difference between these two approaches. Ladd (2008, Table 3.1), for example, shows a potential mapping between British style traditions and AM tone sequences. The question for us is which does a better job of explaining how intonational differences relate to differences in meaning. A closely related question is which phonological elements have distinct contributions to utterance interpretation. We would like to know, for example, whether rises mean something abstracted away from the previous accent.

The AM approach has advantage of having been operationalized for larger scale annotation of English intonation in the form of the ToBI (Tone and Break Indices) labelling scheme (Silverman et al., 1992). Beyond the inventory presented in Pierre-

humbert (1980), this transcription system includes a a downstep diacritic (e.g. !H*) for H type accents that are phonologically lower than previous H targets. The set of pitch accents is reduced to five (L*, H*, L+H*, L*+H, H+!H*). This approach is more or less assumed to be the norm in many analyses of prosody from the theoretical side meaning (cf. (Kadmon, 2001)). Several works have attributed specific meanings to different elements of the ToBI inventory (Pierrehumbert and Hirschberg, 1990, Steedman, 2000). However, this becomes problematic given the uncertain phonological status of some accent types, in particular L+H* (Dilley, 2010). Moreover, other accounts of prosodic meaning still tend to describe contours in an off-ramp style (Gunlogson, 2003, Nilsenova, 2006, Constant, 2007).

While ToBI serves as a convenient short hand for discussing the shape of intonational events, larger scale intonation annotation is a time consuming process. While detection of intonational events (accents and boundaries) obtains quite good inter-annotator agreement, classification of pitch accent type does not fare as well. For example, Syrdal and McGory (2000) reports pairwise agreement on presence or absence of a tone at around 90%, while pairwise agreement was greater than 50% for only 2 of 8 pitch accents types (including downstepped tones), with most confusion being between L+H* and H*. So, a natural question is whether the ToBI tones really tell us what we want to know.

Since ToBI tones are really descriptions of the pitch contour rather than some abstract communicative category, albeit welded to a specific phonological theory, we would like to know whether we can get the same information direct from the signal. A direct modelling approach for prosodic analysis is well motivated from an empirical perspective. Shriberg and Stolcke (2004) show the utility of direct modelling of prosodic features for various speech applications. Similarly quantitative template and target approximation techniques have also proven helpful for understanding the interaction of lexical tone and intonation in Mandarin Chinese (Yuan et al., 2002, Liu and Xu, 2005). The motivation for a direct approach here is particularly salient since methods for developing, for example, automatic speech recognition systems depend on large amounts of training data. In fact, it seems that prosodic studies of ToBI labelled corpora generally are accompanied by a complementary acoustic analysis, e.g. Gravano (2009), Liscombe (2007).

One benefit of starting with a ToBI labelled data set is that intonational events are identified, and so acoustic analysis can be directed at there. However, once regions of interest have been identified it seems plausible that we could characterize the pitch accent or boundary shape using functional techniques from the extracted F0 measurement. Such techniques have already been employed for modelling contour shapes in useful ways. For example, Grabe et al. (2007) show how Legendre polynomial decomposition relates to the high and low tones of ToBI style transcription,

while simultaneously capturing gradient differences in pitch level. Moreover, it seems that a direct modelling approach could be useful in determining whether an off-ramp or on-ramp analysis is more useful for investigations of prosodic meaning.

In general, it seems to really understand how the prosodic information affects interpretation we need to know why we get the pitch accents and boundary tones (or perhaps falls and rises) where we do. Since ToBI annotations just give a representation of intonational *form*, this is not something these sorts of annotations can determine alone. Even with good direct F0 measurements, some perceptual filter might still be necessary in order to understand how those measurements relate to attitudes or structural categories. So, it seems to determine what makes a good input for the prosody-meaning mapping we have to look at what prosody is mapping to. That is, we need some tools to think about what's in a dialogue. This is the topic of the next section.

2.3 Dialogue Structure Basics

A common way to break down what is going on in a dialogue is to think about it as a multiplayer game (Carlson, 1983, Kowtko et al., 1991, Ginzburg, 2012, a.o.). The objective of that game changes from dialogue to dialogue, but the basic board, pieces and rules remain the same. Participants have access to various sources of information during the game which then influence how the dialogue proceeds. We refer to shared or public beliefs as the *common ground*. What a participant does during the dialogue game also depends on their *private beliefs*. We can think of the pieces involved here as different sorts of linguistic objects, e.g. sentence types. Different moves are available to different pieces. For example, declaratives can be interpreted as assertions or questions. What moves are available at any given time depends on pre-conditions of specific moves, the current state of the dialogue (i.e. the state of the board) and, of course, the general rules of the game.

At a high level the assumption is that participants are *co-operative* in the sense of Grice (1975): they tell the truth and they are try to complete the tasks at hand in the most efficient way possible. In this way, the rules of dialogue center around the notion of coherence. Broadly this means that the moves should be interpretable in a way that make senses given the tasks at hand and what has come before. For example, at the local level the assumption is that the questions are followed by answers. To get a grip on this we need to keep track of the questions and tasks that come up and how they are related to one another. The following sections review how these notions have been implemented in several works. Note, however, this is not meant to be exhaustive review, nor is the goal to argue that this is the only way to model

dialogue. Instead, the aim is just to develop the infrastructure upon which we can start discussing the contribution of prosody.

2.3.1 The Common Ground and Public Beliefs

Our basic of idea of how dialogue changes through times is that a dialogue move maps the current context to a new one. Context is usually described in terms of *commitments*. These are in turn modelled as propositions, i.e. sets of possible worlds. I will follow the approaches such as Gunlogson (2008), Farkas and Bruce (2010) and Ginzburg (2012) in separating out the commitments of individuals from the set of commitments shared by all of the dialogue participants, i.e. the *common ground* (Stalnaker, 1978, 2002). Following Ginzburg (2012), we also make the distinction between public commitments from participants' private beliefs.

(1) **Public Commitments** (Public(A,t)): The set of propositions participant A has publicly committed to at time t in the dialogue.[2]

(2) **Private Beliefs** (Private(A,t)): The set of propositions participant A is privately committed to at time t in the dialogue.

(3) **Common ground (cg):** The set of public commitments which are shared by all discourse participants.

A large part of the game of dialogue then is a process of updating discourse commitments. However, what goes into the common ground depends on the evaluation of propositions by individual participants. To do this, participants will need draw on their private beliefs while remaining consistent with their public commitments. We can model this in terms of the notion of *conversational background* developed in Kratzer (1981) to analyze the semantics of modals. Conversational backgrounds are functions from worlds to sets of propositions. These functions pick out sets of propositions that hold in that world that describe different types of information. For example, $f_{\text{epis}(w)}$ picks epistemic knowledge in world w, $f_{\text{deon}(w)}$ picks out the deontic knowledge, i.e. 'the rules', of that world. Of course, there may be a lot of overlap between the conversational backgrounds of participants, and part of the conversational background may actually be about shared beliefs, e.g. 'The sun rises in the

[2]This is the definition of discourse commitment from Farkas and Bruce (2010) except I include propositions that are mutual commitments in this set a la Gunlogson (2008). I think this is a clearer representation of the information in the dialogue, but we could easily revert to Farkas and Bruce's formulation without losing anything.

Level	Clark (1996)	Allwood et al. (1992)	Clarification example
1	Attention	Contact	Did you say something?
2	Identification	Perception	What word did you say?
3	Understanding	Understanding	What did you mean by x?
4	Consideration	Attitudinal reaction	Should I put this into the common ground?

Table 2.1: Levels of Grounding (cf. Benotti (2009))

east'. However, we reserve the term *common ground* for those propositions which have been publicly accepted in the discourse.

One of the things that differentiates dialogue from games like chess or checkers is that dialogue participants don't have an actualized version of the board and pieces that can be objectively inspected. Instead, we only have our private versions of what is happening in the dialogue. So, in reality the common ground is an individual's model of shared commitments. In fact, we can think of conversational backgrounds as a way of picking out a participants (private) view of what's in the common ground.[3] More generally, they give us a way of formalizing how the evaluation of an utterance might change depending on the type of information it is evaluated against.

The corollary of this is that there can be mismatches in what participants view as agreed upon commitments and what relevant evidence a claim should be evaluated against. This motivates the seemingly large amount of linguistic resources that go into clarification (Purver, 2004, Schlangen, 2004, Benotti, 2009). Clarifications can address mismatches at several levels. The influential work of Allwood et al. (1992) and Clark (1996) identify four levels where *grounding* problems can potentially occur. (cf. Table 2.1.[4]

Clarification requests are often triggered by low level problems with the signal, e.g. noise masking a word. However, clarification may also be necessary when a participant simply doesn't believe that a proposed addition is true. It is important to note however, that a speaker's private beliefs can diverge from what they're willing to *accept* in a dialogue (Cohen, 1992, Stalnaker, 2002, Saget et al., 2006). It is not too hard to see that what is acceptable can be influenced by the goals of the dialogue. In cases where high precision of action or co-ordination of beliefs is required. For example, we would expect a higher bar for acceptances in an exam than in a small

[3]In the rest of this dissertation we will mostly be dealing with public commitments, however the notion of conversational background will come up again when we look at the semantics of cue words in Chapter 4.

[4]For Clark (1996), grounding is the establishment of it as 'part of the common ground well enough for current purposed' (pg. 221).

talk scenario. In general, to model what is going on in dialogue we need to keep track of it's task structure.

2.3.2 Tasks and Questions

A major difference in treatments of context is in how much of the general task structure is taken into account. From the semantic/truth evaluation point of view, context is usually thought of in terms of intersections of sets of propositions, i.e. a set of possible worlds. This washes out discourse structural relationships between propositions in the context. However, many accounts of discourse meaning have acknowledged that more is needed to track the dynamic task structure of discourse. This is most commonly formalized in terms of a *Questions Under Discussion* (QUD) stack (Roberts, 1996, Ginzburg, 2012). The idea is that a discourse is driven by questions being pushed on to or popped off of this stack. The stack nature of the QUD can be represented as a tree, where subquestions are questions that have been pushed on to the stack in order to resolve an older related question.

The question at the topic of the stack dictates the current *discourse topic*[5] As such, this current question under discussion has a large part to play in what is deemed relevant. Similarly, the assumption is that a co-operative participant will attempt to resolve this question is a truthful and efficient way. However, to do so they might need to push a subquestion onto the stack as a part of a *strategy* to answer a dominating question (cf. Büring (2003)).[6]

Roberts, following Stalnaker (1978), assumes that 'the primary goal of discourse is communal inquiry the attempt to discover and share with the other interlocutors "the way things are", i.e. to share information about our world'. The assumption from here is that the objective of dialogue participants is to identify one world from the set of possible worlds. While this seems to hold at a very high level, it doesn't quite capture the nature of explicitly goal-oriented dialogues that one would have, for example, talking to a call center operator about paying a bill. In such cases, the primary goal would be, of course, paying the bill. In this sort of task-oriented dialogue, instructions rather than assertions become the order of the day. To deal with this, we make use of the *To-Do list* of Portner (2004). This represents the actions that a particular participant is committed to in a discourse. For Portner, the To-Do

[5]Note: This is not the same as an information structural topic/link in the sense of Vallduví (1990) or Eilam (2011).

[6]Roberts (1996) assumes that questions and subquestions are in an *entailment* relationship in the sense of Groenendijk and Stokhof (1984) for reasons of expository simplicity. This is quite a strong assumption and Roberts she notes that things are actually more complicated. We will see some examples where this is the case in Chapter 6.

list maps agents to properties, which in in turn gives participants a way of ranking possible worlds (from the common ground). The goal is then to make those agents have those properties.

2.3.3 Dialogue Moves and Question Answer Congruence

The idea then is that different sentence types act on different structures: declaratives act on the common ground (adding discourse commitments), questions act on the QUD (setting the discourse topic), and imperatives act on To-Do lists (adding action commitments). However, this view of things only really captures one part of the story, where additions to discourse structures are attempted. That is, the *proffering* side. The other part is about whether or not the proffered content is actually accepted, i.e. the *response* side. This aspect of discourse is explored from the QUD perspective by Farkas and Bruce (2010) who look at discourse semantics of simple polarity responses like *yes* and *no* (cf. Ginzburg (2012)). This brings us to the modelling of the actual dynamics of dialogue in terms of moves.

We can view the framework presented in Farkas and Bruce (2010) as a more proceduralized version of the dialogue dynamics based on the notion of the (Robert's) QUD.[7] Seen in this way, the goal of a dialogue is to empty the QUD stack. For Farkas and Bruce the QUD stack records the 'at-issue' content in a discourse. As in, Ginzburg (2012) both assertions and polar question push a proposition, p say, onto the QUD stack.

The main innovation of Farkas and Bruce's approach is that they assume that proffering moves, e.g. assertions and questions, project future common grounds, *projected sets*, in which the current question under discussion is resolved one way or another. The canonical response is a move which selects one of these projected sets. The difference between assertions and polar questions then lies in (i) what they project and (ii) whether or not a public commitment is made by the speaker. On the one hand, assertions project $\{cg + p\}$ while adding p to the speaker's public commitments. On the other hand, polar questions project two sets $\{cg{+}p, cg{-}p\}$, but don't add p to the speaker's public commitments. In either case the barest canonical response to a positive assertion or polar question is a *yes* $(cg + p)$ or *no* $(cg - p)$. The default is assumed to be the affirmative, under the assumption that denial moves are more marked. Impressionistically, silence can also be taken for acceptance for assertions, but not for questions since the latter don't add any public commitments on their own. We can similarly extend this so that instruction moves, i.e. imperatives,

[7] Farkas and Bruce (2010) rename the QUD stack the *Table*. Since this is more or less a cosmetic change, I will stick with QUD since it is already widely used in the literature.

project future versions of the to-do list which are also mediated through the QUD stack.

In the rest of the dissertation I will basically follow Farkas and Bruce's approach, with the understanding that their projected sets basically recaptures the notion of *Question Answer Congruence* (QAC) often invoked in theories of focus and prominence (Von Stechow, 1991, Krifka, 2001). In the following, we see that it is infelicitous to put main sentence stress on the subject responding to an object wh-question.

(4) Q: What did Fred eat?
   a. A: Fred ate the BEANS
   b. #A': FRED ate the beans.

Büring (2008) formulates this in terms of Rooth (1985) style focus semantics. Let $[\![U]\!]^o$ and $[\![U]\!]^f$ be the ordinary and the focus semantic values of the utterance U respectively. The focus semantic value adds a variable for each F-marked unit of the utterance.[8]

(5) Question-Answer Congruence (QAC): A is a felicitous answer to Q only if:
   a. $[\![Q]\!]^o \subseteq [\![A]\!]^f$
   b. There is no alternative focusing A' of A which has less F-markings and meets (a).

For a polar question $?p$ the ordinary semantic value $[\![?p]\!]^o$ is just $\{p, \neg p\}$, as in Farkas and Bruce's projected sets, so a *yes* or *no* will suffice. For the wh-questions things are more complicated, since the focus semantic value of the answer $[\![A]\!]^f$, depends on where the main sentence prominence falls.

(6) $[\![Q$ = What did Fred eat?$]\!]^o$ = {Fred ate x | $x \in D_e$}
   {Fred at the beans, Fred at the broccoli,...}
   a. [Fred ate the BEANS]^f = {Fred ate x | $x \in D_e$}
   b. [FRED ate the beans]^f = {x ate the beans | $x \in D_e$} $\not\subseteq [\![Q]\!]^o$

We will see later that QAC has a large part to play in the relationship between prosody and information structure, and so will be crucial for understanding how terminal rises are interpreted. The key point for now is that we can characterize dialogue moves in terms of what they do to various discourse components: public

[8]For the current purposes, F-marked means marked as contrastive. In reality, this is often but not always signalled through pitch accenting. We'll come back to the issue of focus and contrast in Chapter 5.

commitments, common ground, the QUD stack, and the To-Do list. Moreover, we can use QAC to characterize what a canonical (*direct*) response to a proffering move is.

2.3.4 Dialogue Act Schemas

So far, however, we have basically treated move type and sentence type as the same thing. The approaches cited above only really deal with declaratives assertions and syntactic polar interrogatives (and occasionally imperatives). As such we have only considered a very few type of dialogue move. Any substantial look into characterizing the moves of real dialogue quickly will tell us that broad the assertions, question and instruction categories are not enough to obtain good coverage of the data.

Larger dialogue move schemas have arisen to help with the development of automated dialogue systems where wide coverage is more important. A widely used labelling schema of this type of is DAMSL (Dialogue Act Markup in Several Layers, Allen and Core, 1997). Core and Allen (1997) define dialogue acts (communicative actions) as 'explicit manipulation of the common ground'. Dialogue acts are generally seen as an extension of Searle's (1969) speech acts. However rather than attributing a single communicative function to an utterance, they encode several interpretational layers. They note that, for example, responses can also be informing. In general, utterances that are about speaker's reactions to have a *backward communicative function* (e.g. agreements, answers), while non-reactive utterances have *forward communicative function* (e.g. assert, information-request, action-directive). This basically corresponds to the proffering/response distinction discussed above. DAMSL also distinguishes moves types on several other levels. For example, statements which attempt to 'affect the beliefs of the hearer' (Assert moves) get different labels to repeats and acknowledgements. Similarly, a response could accept a proposal besides being an answer to a question. The DAMSL scheme also distinguishes talk directed at achieving the task at hand from talk about the dialogue itself.

While DAMSL was intended to be used across different dialogue types markup system, the types of dialogue act that are used in actual annotation projects still need to be modified for different domains. Jurafsky et al. (1997) present a modification of DAMSL (SWBD-DAMSL) for the Switchboard corpus of conversational telephone speech (Godfrey et al., 1992). The Switchboard corpus differs from the TRAINS corpus that DAMSL was initially tested on in capturing small-talk type conversations rather than goal directed dialogues, e.g. solving shipping problems, of the latter corpus. The initial SWBD-DAMSL tag set included 220 unique tags representing different communicative actions for 205,000 utterances. These tags were clustered

Tag	Example	Count	%
Statement	Me, Im in the legal department.	72,824	36%
Backchannel	Uh-huh.	37,096	19%
Opinion	I think its great	25,197	13%
Agree/Accept	Thats exactly it.	10,820	5%
Abandoned/Turn-Exit	So, -/	10,569	5%
Appreciation	I can imagine.	4,633	2%
Yes-No-Question	Do you have to have any special training?	4,624	2%
Non-verbal	<Laughter>, <Throat Clearing>	3,548	2%
Yes answers	Yes.	2,934	1%
Conventional-closing	Well, its been nice talking to you.	2,486	1%
Uninterpretable	But, uh, yeah	2,158	1%
Wh-Question	Well, how old are you?	1,911	1%
No answers	No.	1,340	1%
Response Ack	Oh, okay.	1,277	1%
Hedge	I dont know if I'm making any sense or not.	1,182	1%
Declarative Question	So you can afford to get a house?	1,174	1%
Other	Well give me a break, you know.	1,074	1%
Backchannel-Question	Is that right?	1,019	1%

Table 2.2: Most frequent SWBD-DAMSL labels, from Jurafsky et al. (1997)

Move	Percentage
Acknowledge	20%
Instruct	16%
Reply-y	12%
Check	8%
Explain	8%
Ready	7%
Align	7%
Query-yn	7%
Clarify	5%
Query-w	3%
Reply-n	3%
Reply-w	4%

Table 2.3: HCRC Map Task move percentages from Kowtko (1997), $N = 25,945$.

into 42 tags for reasons of data sparsity. Table 2.2 shows the most frequent SWBD-DAMSL labels in the Switchboard annotation. Additions for the SWBD-DAMSL set were categories like Opinion (forwards) and Appreciation (backwards). Several types of question were also added (e.g. declarative, tag).

Other annotation efforts have used much smaller DA sets. For example, the SimpleMDE annotation of Switchboard uses only four labels: statement, question, backchannel and incomplete. Differences in tag sets developed for other corpora outside of the DAMSL fold highlight the different requirements of different speech genres. The AMI meeting corpus tagset includes 15 tags falling into six groups: information exchange (Inform, Elicit Inform), possible actions (Suggest, Offer, Elicit Offer Or Suggestion), comments on previous discussion (Assess, Elicit Assessment, Comment About Understanding, Elicit Comment About Understanding), social acts (Be Positive, Be Negative), special completion classes (Backchannel, Stall, Fragment), and the bucket Other label (Renals et al., 2007). Like Switchboard, the types of interaction in the AMI corpus are to do with information exchange and coming to consensus. There is a conspicuous lack of instruction type moves which are much more prevalent in, for example, the HCRC Map task annotations where the goal is for one participant to get the other to reproduce a route on a map (Anderson et al., 1991).[9]

The Map Task move set includes six initiating moves (instruct, explain, check, align, query-yn, query-w), five response moves (acknowledge, reply-y, reply-n, reply-w, clarify), and a pre-initiating move (ready) (Carletta et al., 1997). Proportions of

[9]We'll discuss the map task in further detail in Chapter 7.

move types in the HCRC Map Task corpus are shown in Table 2.3. Each initiating move marks the beginning of a specific type of conversational game which projects a specific sort of follow up (Kowtko et al., 1991). For example, Instructing games finish with an acknowledgement indicating an action has been accepted. Games can also be nested or broken off.

The idea of forward/backward orientation and moves initiating conversational games is very similar to the approach of, for example, Farkas and Bruce (2010) discussed above. What's missing from the dialogue act centered studies is the notion of the QUD which centralizes the dialogue dynamics and projects discourse expectations about relevance, for example. In DAMSL, for example, while some moves are characterized as backward looking, there generally no one move type that is distinguished as a canonical response for a forward looking move (although we can determine whether some are more probable than others). In fact, the goal of data driven studies that make use of such dialogue act schemas is often to infer higher level structure using statistical or machine learning methods for the purposes of improving act or word recognition (Shriberg et al., 1998, Wright Hastie et al., 2002, Rangarajan Sridhar et al., 2009).

The lack of an explicit common ground gatekeeper like the QUD makes such bottom up work in some ways more similar to relational approaches to modelling discourse such as SDRT (Segmented Discourse Representation Theory, Asher and Lascarides, 2003). In this framework utterances in are connected by a rich set of rhetorical relations such as *elaboration* or *explanation* which give a structured representation of the context. The discourse structure is built incrementally and the current structure determines possible attachment points. A glue logic, with defeasible rules, determines where an utterance is attached in the structure and what relation is used.[10] Thus, rather than a discourse tree (via the stack) we get a discourse graph. So, discourse expectations come from information dispersed over the entire relational structure. Like the dialogue act annotation schemas mentioned above the set of identified rhetorical relations is very large and oriented more towards function rather than form. However, the emphasis here is more on relationships between utterances rather than communicative actions. We can again contrast this with QUD based frameworks like Farkas and Bruce (2010) which tend to deal with moves associates with particular sentence types and where the relationships between public commitments don't play much of a role.

[10]Defeasible rules are of the form: P > Q, "if P then normally Q".

2.3.5 Where does prosody act?

Of course, it is entirely possible that a full model of dialogue requires both a relational structure and a more centralized QUD driven mechanism for managing notions like public commitments. The question for us is then how and where does prosody enter the interpretational mix? There are several potential points at which rises could have an effect. At the lowest level we view rises as providing some sort of meaning directly to the utterance semantics, like a propositional attitude. Similarly, we might expect rises to have something to do with distinguishing move types. This would reflect the intuitive association between rises and questions of various syntactic forms. On the other hand, intonational components may act on the discourse structures at a higher level rather than on their carrier utterance. For example, rises might place a post-condition on the QUD. Similarly, it might affect where an utterance can enter the relational structure of discourse. It might also be the case that rises directly signal an attitude or affect that is independent of semantic/pragmatic and structural considerations attached to an utterance. In the next section we look at analyses that span this large range of views.

2.4 Previous Analyses of Final Rises

2.4.1 Forward Dependence and Non-Finality

The starting point for most analyses of terminal rises is the observation that statements generally fall while questions rise (O'Connor and Arnold, 1961, Halliday, 1970). However, we need to break this down in terms of sentence type and move type. In general, it seems the default form for polar questions with auxiliary inversion (in American English) and declarative questions is a final rise, while wh-questions exhibit a final fall. Haan (2002) has argues that the less an utterance is syntactically marked as a question, the more intonationally marked it will be (her *functional hypothesis*): wh-questions are the most syntactically marked for questionhood, declarative questions the least. However, this characterization is not rigid. For example, Haan found that 64% of wh-questions were produced with a final rise in a Dutch reading task. Similarly, corpus studies have found that not all declarative and polar questions rise (Brown et al., 1980, Geluykens, 1988, Hirschberg, 2000, a.o. for English). We might expect that some of that variability comes from the fact that questions with the same syntactic form still have different interpretations and discourse uses. This is particularly salient for elided questions. For example, elided wh-questions can be used to seek information (e.g. *What (did you do)?*) or to elicit repetitions (e.g. *What*

(did you just say)?). So, simple marking of sentence or move type does not seem to be enough to account for the distribution of rises.

The terminal rises that are associated with these sorts of questions are canonically nuclear rises, i.e. L*+H H-H% in ToBI. This begs the question of whether these rises should be treated in the same way as rises in fall-rise contours (e.g. H*L-H%). Unsurprisingly, analyses of prosodic meaning vary in this respect, and more generally in what parts of the assumed phonological inventory they assign meaning to.

Pierrehumbert and Hirschberg (1990, P&H) attempt to assign distinct meanings to all elements of the AM/ToBI inventory (except downstep), which are assumed to be independent of prosodic context. Pitch accents mark variables in an open expression:

(7) GEORGE likes CAKE
   a. P(x,y) = likes(x, y)
   b. x = George, y=cake

H* accents mark units that are to be predicated, i.e. added to the common ground, while L* accents mark units that aren't to be added. This covers the observations that questions don't usually add public commitments and are usually produced with nuclear rises. However, P&H also suggest a reason for non-predication might be that the informational unit is already in the common ground. Bitonal accents additionally evoke a scale, but keep the predication type of starred tone. Edge tones indicate whether the utterance should be interpreted with respect to subsequent material. For the boundary tones, H% indicates that the utterance that is the case, while for L% this is left unspecified (the phrase accents get a similar interpretation). So, H% signals some sort of *forward dependence* in the sense that the H% is in a hierarchical relationship between the intentions of the current utterance and one to come (cf. Grosz and Sidner (1986)). In the following, the idea is that the first two sentences are interpreted as evidence supporting the last.

(8) a. George likes cake$_{LH\%}$ Pierrehumbert and Hirschberg (1990, (61))
   b. He adores pie$_{LH\%}$
   c. He'll eat anything that's sweet and calorific$_{LL\%}$

This presents a unified analysis of the rise in questions (H-H%) and continuations (L-H%). The fact that H% does not make a good discourse end is treated as a corollary. While we could take this to mean the L% is something of a vanilla default, P&H do suggest that it has a segmenting sort of meaning. As pointed out by Hobbs (1990), since dominance is allowed to go both ways, the meaning of H% is very weak. Moreover, P&H allow satisfaction-precedence relationships to license rises.

This basically means that H% is licensed as long is it is in some sort of intentional relationship with something that comes later. This should be the case for any non-final utterance in a discourse that maintains relevance (Banuazizi and Creswell, 1999).

Hobbs (1990) suggests that this problem could be obviated if H% were analyzed with respect to a richer theory of discourse structure (since Grosz and Sidner's approach only has dominance and satisfaction-precedence). More specifically, he suggests a theory of discourse structure that uses builds hierarchical discourse structure in a recursive fashion, so that sister nodes represents discourse segments in a particular relation. In this way, H% is taken to signal discourse *openness* or *non-finality*: we expect that another segment needs to join with the H% bearing one to form a larger coherent segment. This sort of non-finality type analysis has also been advocated for in Cruttenden (1981) and Bartels (1999). Similarly, Brown et al. (1980) note that 'Not-low' terminals are associated with there being more to come on the current topic. In the QUD driven dialogue framework outlined above, the obvious way to model this non-finality would be to say that rises signal that some salient question under discussion (or task) is unresolved. From this point of view, rises point up into the discourse structure rather than forward. This approach has the benefit of giving us more of a notion about what the dependency might be and so how the discourse can be closed.

Like P&H, (Bartels, 1999) assigns discourse meanings to all of the elements of the ToBI inventory. Most significantly, she argues that the phrase accent, L-, adds an [assert] morpheme, which explains why fall-rises are still assertive while being terminally rising. To extend this to default falling wh-questions, she needs to assume that these sorts of questions assert via an existential presupposition, i.e. *Who ate the cake?* presupposes that someone ate the cake.[11] However, it is not clear how this assertion morpheme works in varieties of English for which falling polar questions are unmarked, e.g. Standard Southern British (Grabe, 2004).

Taking a broader look at the English usage also brings up problems for the compositional meanings proposed in in P&H. In particular, the idea that L* denies predication runs into trouble with the data of McLemore (1991) who analyzes rising assertions, i.e. *up-talk*, as having a rising (L*H%) nuclear tune in spontaneous information giving monologues (so L*H% marks new information). So, we would want to see if there is really a case to be made that these intonational elements contribute these compositional discourse meanings separate from what is present in the utterance semantics and the larger discourse context. Similarly, we would also like to know whether the meaning of these contours is what is predicted by the composition of these elements.

[11] It's a bit far a field to go into now, so I shall simply note that the presence of the presupposition is disputed (Eilam, 2011).

2.4.2 Commitments and Contingency

The common thread through these accounts is that rises say something about how an utterance relates to the rest of the discourse structure. I have suggested that the relevant part of the structure is the QUD. However, a different way of looking at this relationship is to cast it in terms of speaker and hearer responsibilities, i.e. public commitments. This is the approach pursued in Steedman (2000) and updated in Steedman (2007). Like P&H, Steedman takes a strongly compositional view of intonational meaning. However, this account is founded on a more articulated idea of the information structure of an utterance. Here, pitch accents are assumed to mark informational units: both H* and L* mark rhemes while bitonal accents mark themes.[12] H* type accents, bitonal or not, indicate agreed upon informational units, while L* accents mark contentious units.

For the boundary tones, H% mark units that the speaker claims the hearer is commitment to, while L% mark units that speaker is committed to. This once again, draws mainly on the association between questions and rises. Like Bartels (1999), to explain why wh-questions fall, the account assumes that the speaker has responsibility for such questions through an existential presupposition. Steedman further argues that forward dependency or continuation type aspects of final rises are in fact conversational implicatures from this signal of hearer commitment. By putting the responsibility for utterance into the hands on the hearer, the speaker implicitly signals that more needs to be said. However, it's seems equally possible that the reverse is the case: the hearer is implicated as responsible when as an epiphenomenon of the speaker signalling the QUD is open, assuming that being co-operative the speaker would have resolved the issue if they could have. Steedman suggests, for example, that lists of instructions sound kinder and less discouraging with final (low) rises, because they suggest that the hearer is in control of the action. Besides the fact that not all rise linked lists sound kind, this doesn't cover floor holding situations like the following where it attributing hearer responsibility seems unnecessary. We can, however, easily understand the presence of the rise as signalling that Bert has not finished listing all of the tasks he has to complete.

(9) a. A: Hey Bert, what do you have to do before you leave?
b. B: Well, let's see...I've got to deposit my dissertation$_{\text{L-H\%}}$
c. B: I've got to find a new tenant$_{\text{L-H\%}}$
d. B: I have to sell all my furniture$_{\text{L-H\%}}$
e. B: Fit all of my stuff into two suitcases$_{\text{L-H\%}}$

[12] Rhemes roughly correspond to foci, i.e. the part that answers the immediately dominating QUD; themes correspond to the rest of the utterance, which includes the topic, cf. Vallduví (1990). We'll come back to this in Chapter 5.

f. B: And I've got to visit all my relatives
g. A: Ok, I'll leave you to it!

This hearer responsibility approach is central to the account of declarative questions in Gunlogson (2003). The general idea is that declaratives always assert in the sense that they mark an attempt to add something to the common ground. However, for rising declaratives the responsibility for that addition is given to the hearer. This gives them the aura of a question, while explaining why declarative questions seem to express a bias that positive polar interrogatives generally do not. This bias is brought out in, for example, interview situations, where the rising declarative seems to assume that the addressee is a communist.

(10) a. You're a communist?
b. Are you a communist?

The licitness declarative questions under this view was determined by a contextual bias condition: that the addressee was already committed to the propositional content of the declarative (though not necessarily publicly). This dovetails nicely with the hearer commitment analysis, since then the rise just reflects this perceived/inferred hearer bias. Gunlogson (2008) presents a revision of this account based on an analysis of initiating declarative questions. In many such cases, previous commitment of the addressee does not seem to be necessary, as in the following.

(11) At Tim's graduation, where Tim is standing next to an older woman. Jack to Sophie:
a. Are you Tim's mother?
b. You're Tim's mother?
c. [So/I guess] You're Tim's mother.

To cover this data, Gunlogson reanalyzes rises as expressing discourse contingency:

(12) A discourse move m by an agent A is contingent upon a discourse condition d if:
a. d does not obtain at the time of m
b. It is inferable in the discourse context that the update effected by m is to be retained only if d obtains after the discourse move immediately succeeding m.

In this account, Gunlogson formalizes the idea that commitments needs to have sources. When a speaker wants to commit something to common ground but doesn't

have the authority to do it, they may ask the other participant to ratify the commitment. Thus, adding a rise to a declarative indicates a contingent commitment.

(13) A discourse move m committing an agent A to p is contingent upon ratification by an agent B, A $\neq$ B, if:

a. B is implicitly authoritative with respect to p at the time of m b. It is inferable in the discourse context that A's commitment to p will be withdrawn unless the discourse move immediately succeeding m has the effect of committing B to A as a source.

This notion that a move depends on something to come is stronger than a non-finality type analysis. In fact, it can be seen as a strong version the forward dependency account of Pierrehumbert and Hirschberg (1990). In this case, the move bearing the rise, rather than some higher level relation, is not realized until further action is taken by the addressee. It is important to note here, however, that Gunlogson only deals wit nuclear rises on declaratives. It's not clear whether such a strong post-condition fits the interpretation of, for example, fall-rise declaratives. However, it could be that a more general version of the contingency notion will be helpful in explaining how rises fit in with different dialogue contexts.

2.4.3 Tests, Modals, and the Common Ground

The accounts reviewed so far all cast rises in discourse structural terms, i.e. above semantics. However, some accounts have attempted to formalize the contribution of rises in a more traditional semantic light. In Nilsenova (2006), rising intonation is treated as an intonational adverb expressing uncertainty similar to the modal *might* in update semantics (Veltman, 1996). That is, $\diamond p$ *tests* if there is at least one possible world in the common ground where the propositional content of the declarative is true. As such, this something of a formal implementation of the analysis of accents in Gussenhoven (1984), in which nuclear rises signal a test on a variable and the common ground. It is also related to the analysis Beyssade and Marandin (2007), who analyze rises in French as signalling some sort of defectivity in the common ground in a Ginzburg (2012) type framework. This view differs from Gunlogson (2003) and Gunlogson (2008) in that it does not propose actual addition or change in the common ground or the participants public commitments. Instead, the informational redundancy of this test leads to the impression of bias and questioning aspects are derived as by-products of Gricean Conversational Maxims. Nilsenova claims that final rises on polar interrogatives indicate that the questioner is willing to accept uncertain answers like 'possibly'. Conversely, more biased questions are likely to fall since the

questioner expects that the answer is known. Nilsenova also suggests that as the content of a rising declarative is already in hearers beliefs, it signals that updating has not been performed successfully.

The corollary of this analysis is that rises express a propositional attitude: epistemic uncertainty. If the speaker were certain after all, they would simply make an assertion, rather than setting a test. Like Gunlogson (2003), this analysis puts a strong post-condition on the discourse – the test should be resolved. It also excludes continuation type rises from the analysis. Moreover, Nilsenova claims that intonation should be treated in the semantic system. So it is not clear how robust this account is to variation in the data, such as the observation that declaratives can act as questions while exhibiting final falls. In this case, it is not clear whether such questions could obtain the same semantics as rising declaratives. This is fine, however, for a Gunlogson style analysis since the level rises are operating at is above the semantics, hence other contextual factors can contribute to the understanding of contingency. Furthermore, it is not clear that uncertainty is always the appropriate attitude to associate with rises.

2.4.4 Speaker Attitudes and Affect

Earlier impressionistic studies of intonation often often associated with specific affects or attitudes (Palmer, 1922, Kingdon, 1958, O'Connor and Arnold, 1961). This is characteristic of descriptions of intonation developed to aid learners of English by giving them general rules to follow for particular settings.[13] For example, low falls indicate lack of interest, while high falls makes the speaker seem more involved (O'Connor and Arnold, 1961). Cruttenden (1997) glosses fall-rise contours as expressing 'reservations'. For such works, contours are treated holistically and intonational meaning varies greatly with semantic content and sentence type. As such, strong generalizations about the relationship between intonational components and meanings do not abound. In this vein, Bolinger (1989) notes that that there is no a priori limitation of the type of pitch contour a statement can have. However, he suggests that certain speech acts tend to be marked intonationally. For example, he claims that pronouncements tend to be falling, while observations often exhibit a fall-rise. Bolinger characterizes the latter as pointing to something obvious, hence not informing, as well as projecting casualness.

The attitudes most often associated with final rises in instrumental studies are uncertainty (Gravano et al., 2008, Pon-Barry, 2008, Litman et al., 2009) and politeness/pleasantness (Uldall, 1962). Gravano et al. (2008), for example, found that

[13]Hence, also the focus s on mapping contours to sentence types.

resynthesized downstepped contours make a utterance sound more certain than the 'usual' declarative while rising contours sounded uncertain. The understanding here is that questions inherently convey some amount of uncertainty in that they request information, rather than proffering it. However, Brown et al. (1980) notes that rises can be associated with both kindness and hectoring depending on the situation and voice quality. Similarly, Scherer et al. (1984) found final falls to be interpreted 'challenging' on polar questions, but not on wh-questions, indicating the the need to take into account textual considerations when attributing attitudinal meaning to intonational parts (cf. Ladd et al. (1985)).

While this sort of direct to attitude type analysis has taken a back seat since the the advent of the AM framework, the connection between intonation and attitude still pops up in recent analyses of rises. As mentioned previously, in analyzing nuclear rises as contributing a something like a possibility modal, Nilsenova (2006) implicitly encodes epistemic uncertainty into the interpretation of the rise. Similarly, Reese (2007) assumes that final rises indicate uncertainty at the level of cognitive modelling. As such, it directs how new information is placed in the rhetorical structure of the discourse, a la SDRT (Asher and Lascarides, 2003), rather than rather acting directly on a model of the common ground. Similarly, Merin and Bartels (1997) analyzes rises as signalling submissiveness and falls dominance in dialogue. The idea is that, for rising declaratives, giving up dominance prevents assertion by the speaker. On the other hand, this analysis assumes that falls assert. This is problematic, given that we know declarative questions can be falling.

Analyses like that presented in Merin and Bartels (1997) are very much inline with the hearer responsibility analyses of Steedman (2000) and Gunlogson (2008). In general, it is not hard to see how an implication of uncertainty could be derived from an analysis of rises as passing the buck to other conversational participants or putting an test on the common ground. However, the idea that rises signal uncertainty is somewhat at odds with recent findings on the prosody of affirmatives, e.g. *okay*, *uh-huh*, in task-oriented dialogue. Rises have been found to be indicative of backchannels in several corpus studies (Hockey, 1993, Savino, 2011, Heldner et al., 2008, Caspers, 2000, Benus et al., 2007). These particles, e.g. *uh-huh*, *okay*, are viewed as signals of attention veering towards an agreement or acceptance interpretation, i.e. the opposite of uncertainty. While a submissiveness analysis might be plausible here, other studies have found socially dominant speakers to produce more rises depending on the situation (McLemore, 1991, Cheng and Warren, 2005). A looming question is whether these attitudinal effects can really be attributed to the meaning of the contour components, or whether they arise from some other channel.

2.4.5 Iconicity and Universality

Part of the problem of dealing assigning meaning to prosodic patterns is that prosodic features seem to signal meaning paralinguistically. So, the question is whether we can treat intonational forms like morphemes in the grammar, or whether we should consider them to be in a separate interpretational channel. Several studies have argued that intonation has special status. The idea is that, while intonational forms can be grammaticalized, there are aspects of intonational interpretation which arise from a pre-linguistic connection between pitch gestures and meaning. For example, Liberman (1979) suggests that 'a substantial portion of the content of the intonational lexicon of English is determined by the universal (better: metaphorical) value of tones and tone-sequences' (pg. 138). Nevertheless, Liberman also argues that a substantial amount of conventionalization 'must be built into the system'. The offshoot of this that even though the phonological system may deal with level tones, rises and falls as gestures may still mean more than their constituent parts. Such gestures are taken to be ideophonic, e.g. echoic words in English like *clang*. These words are strongly influenced by universal considerations – gestures have fundamentally iconic meanings. Nevertheless conventionalization is possible in a certain situations, so that these words/gestures can take on more arbitrary meanings.

This sort of view is explored in a series of works by Gussenhoven (2004) who identifies three *biological codes* exemplifying iconic properties of pitch. The *frequency code*, originating with Ohala (1984), interprets higher voices as less threatening (i.e. shorter larynx, smaller body) and hence more polite, friendly, or submissive. This code is associated with the questioning effect of final rises. The *effort code* suggests that important information is produced with more articulatory effort. This code has been primarily associated with pitch range, or more specifically the 'wider excursion of pitch movement' (Gussenhoven and Rietveld, 2000). This connects production to speaker affect, e.g. the expression of surprise. Conversely, compressed pitch range is linked to negative attitude or boredom. These two codes are *affective* in that they reflect speaker state. The **production code**, however, is more structural. The generalization here is that high beginnings signal new topics while high endings signal continuations. This derives from the fact that energy diminishes during exhalation. New discourse segments align with high pitch, presumably because they come after a pause in which the speaker takes a breath.

Gussenhoven allows for intonational forms to be grammaticalized. However, the expectation is that a listener who does not know the language being spoken will interpret pitch production in this way. In support of this Gussenhoven and Chen (2000a) show that Chinese, Dutch and Hungarian listeners selected (resynthesized) utterances with higher peaks, later peaks and higher final rises as being more question like when presented with utterances from a language unknown to them, even

though each of these languages express interrogativity differently in both syntactic and prosodic terms. However, when it comes to actual analysis of languages this theory becomes very weak. In particular, it is not clear how we reconcile high pitch associated with effort and hight pitch associated with submissiveness, nor does it explain the preference for wh-questions that fall or the fact rising questions can be used as imperatives.

In general, it is not straightforward to tease apart the relative contribution of each code. For example, since high frequency signals submissiveness we might also hypothesize that higher pitch level should signal acceptance of the other speaker's opinion utterance in a backchannel like way. However, this seems to interact with the contribution of the effort code. Moreover, there are various trading relations in play. Gussenhoven (2002) notes a trading relation between peak delay and peak height in terms of prominence (although both can happen together). Similarly, higher register can substitute for wide pitch span. However, language specific differences are evident (Rietveld et al., 1999, Gussenhoven and Chen, 2000b). So, findings related to these codes must be interpreted relative to different thresholds for different languages as well as individual variation.

Hirschberg (2002b) attempts to translate Gussenhoven's biological codes into Gricean style maxims or communicative conventions (Grice, 1975). This takes pitch gestures out of the realm of iconic and into that of co-operation. As such it makes a rather stronger claim than what is conveyed by the biological codes.

(14) Hirschberg's intonational maxims

a. **Maxim of Pitch:** Try to make the rise or fall in the pitch of your utterances correspond to the degree of confidence you wish to convey. Let your pitch rise to convey uncertainty and fall to convey certainty.
b. **Maxim of Emphasis:** Try to make informationally important portions of your speech intonationally prominent.
c. **Maxim of Range:** Let the width of your pitch range reflect the location of your utterance in the topic structure of the discourse. Increase your range to start new topics. Decrease your range to end old ones.
d. **Maxim of Phrasing:** Phrase your utterance so that it is divided into meaningful portions of speech.

Gricean maxims are supposed to reflect aspects of rational co-operative behaviour that apply outside of the language module. However, it is not clear that the prosodic generalizations listed above really form a basis for co-operative behaviour. Nilsenova (2006), for example, notes that unlike Gricean maxims like Relevance, violating these pitch rules don't appear to create implicatures. Instead violating expectations seems

to trigger something more like presupposition accommodation. Similarly, it's not clear what flouting the Maxim of Phrasing would mean besides signalling that a speaker is disfluent. The problem seems to be that this sort of formulation assumes that what were originally envisioned as iconic meanings are the expected interpretation of the pitch gestures even when we take into account language specific conventionalizations. This doesn't seem to be the case for the prosody versions. In fact, it is not at all clear that rises always signal uncertainty even if we just confine our studies to English.

In general, Hirschberg's observations do appear to pick up on some norms of speech but it doesn't seem correct to try to put these on the same level as the usual Gricean maxims. It seems more plausible to interpret the biological codes as a very basic prior bias which can be overruled by other considerations. The question is, how much are these pre-linguistic expectations affected by grammatical and contextual considerations. The sociolinguistic study of uptalk in sorority speech of McLemore (1991) would suggest that the answer is quite a lot. Based on a a detailed qualitative analysis of spontaneous speech, McLemore argues that boundary tunes have a *diagrammatically iconic* meaning: rises are connecting, level tones are continuing, while falls are segmenting. However the actual interpretation of rises depends heavily of situational and cultural conventions. For example, socially dominant sorority members used rises to take and hold the floor in monologues, while rises were perceived as expressing uncertainty in narratives by uninitiated members of the group. So, the social status and the discourse type seem to heavily effect how rises were interpreted.

2.5 The Way Forward

The most salient difference amongst the treatment of rises reviewed above is between the older style accounts that map contours holistically to attitudes, and accounts which look at rises as acting or commenting on some part of the discourse structure. From the former camp, we saw that rises have often been associated with uncertainty or some kind of speaker deference which is sometimes interpreted as politeness. However, perceptions of these attitudes seem easily influenced by non-prosodic features of the utterance such as sentence type.

Table 2.4 gives a brief summary of various analyses discourse structure based analyses of final rises of this kind. This again highlights the division between off-ramp ('rise') and on-ramp analyses (H%). For the former, rises are generally treated as contours which are convex through the nuclear accent of the utterance. Some of these accounts also offer separate analyses of the fall-rise contour. For example, Gussenhoven (1984) analyses fall-rises as selecting a variable in the background. Several analyses have also focused on the differences between falls and fall-rises (Jackendoff,

Analysis	Fall	Rise
Cruttenden (1981)	closed	open
Gussenhoven (1984)	addition	testing
McLemore (1991)	segmenting	connecting
Merin and Bartels (1997)	appropriates dominance	alienates dominance
Gunlogson (2003)	speaker commitment	hearer commitment
Nilsenova (2006)	-	test/uncertain
Beyssade and Marandin (2007)	compatibility	defectivity
Gunlogson (2008)	-	contingency
	L%	H%
Pierrehumbert and Hirschberg (1990)	-	forward-dependency
Bartels (1999)	-	non-finality
Steedman (2000)	speaker commitment	hearer commitment
Reese (2007)	-	cognitive uncertainty

Table 2.4: Summary of rise analyses.

1974, Ward and Hirschberg, 1985, Büring, 2003, a.o.). So, it seems one of the key questions in determining rise meaning is whether the rise in a fall-rise is the same as a nuclear rise in interpretational terms.

Besides the differences in prosodic representation, we can also distinguish these analysis in terms of what part of the discourse structure they appear to act on. Steedman (2000), for example, assumes that rises act on hearer commitments, while Nilsenova (2006) assumes that rises act on the common ground. In a similar vein, I characterize non-finality/forward dependence type analyses in terms of QUD openness. Analyses of the latter type are the most broadly applicable in that they aim to account for all boundary rises, not just nuclear rising tunes. They also are the most firmly situated at the discourse level: they describe licensing conditions for rises in terms of a discourse configuration. This differs from accounts like that of Nilsenova (2006) and Gunlogson (2008) which assume that rises mainly signal something about their carrier utterance. Both of these accounts place something like a test on the propositional content of rising utterance. The offshoot of these accounts is that nuclear rises put a strong post-condition on the dialogue, one that is inherently associated with uncertainty.

This leaves us with several possible views of the prosody-meaning map. To figure out which of these approaches best fits data, we need to answer the following questions.

- What is an appropriate representation for understanding how intonation contributes to dialogue interpretation?

  - Should we treat rises as contributing something separate from pitch accents?
  - Do differences in meaning map better to dynamic contour shapes or tonal targets?
  - Can we avoid manual categorization and model intonational forms directly from the signal?

- What aspect of discourse structure do rises act on?
  - Should we represent rise meaning in terms of a semantic operator that acts on it's carrier proposition or in terms of discourse configurations?
  - What are the pre- and post-conditions of rises?

- How do we reconcile discourse structural approaches with instrumental and impressionistic findings on rises and speaker attitudes?
  - Do analyses focusing on nuclear rises and polar questions/declaratives generalize to other sentence/move types?
  - Under what conditions do rises signal uncertainty or submissiveness? Is this signalled via a separate, possibly paralinguistic, channel to the proffered content?

- What sort of context do we need to take into account in order to interpret rises correctly and understand their distribution?

In the rest of the dissertation, I will attempt to address these questions through a detailed study of how rise interpretation and distribution is affect by (i) utterance semantics, (ii) local context as mediated by the QUD stack, and (iii) higher level features such as task, role and dialect. Moreover, the idea is to use the types of quantitative methods to gain some notion of the variance in the prosodic data – something usually lacking in theoretical accounts of intonational meaning. We also want to be able explain previous findings from other empirical studies. These findings are usually reported as stylized facts, such results really describe characteristics of distributions of acoustic features with respect to discourse or attitudinal categories. It's not clear what the best way is to square this gradience with the categorical nature of theoretical analyses of meaning and prosodic representation, particularly when it comes to perceptual judgements about attitudes. However, it seems that something like the effort code could help us out here. To be able resolve these issues we need to look at something that occurs with some frequency in real dialogue, but which we can also make some predictions about from the theoretical standpoint. As such, our first stop will be cue words.

Chapter 3

Cue Words, Dialogue Acts and Prosody

3.1 Introduction

The general question of this dissertation is how does prosody affect the interpretation of an utterance in a dialogue. To make some headway into this problem, we need to consider sets of data which are tractable in terms of investigating (i) their phonetic detail, (ii) their semantic/pragmatic contribution, and (iii) the interaction between the two. Preferably, we want to consider linguistic objects that are consistently and frequently used to shape the structure of the dialogue. A candidate data set that fulfills these criteria are cue words responses. Cue words include affirmatives, backchannels and short questions like *okay*, *yeah*, *right* and *really*. The discourse markers appear frequently in spoken dialogue and the fact that they have specific segmental forms makes it easier to study the effects of prosody on their interpretation. Moreover, understanding these sorts of markers is important because they indicate both when things are going well and when things are going badly in a dialogue.

The primary role of cue word responses seems to be to evaluate the utterance (or situation) they are responding to, and so guide what is to come next. Consider, for example, the following excerpt based on a dialogue from the Switchboard corpus of conversational speech (Godfrey et al., 1992).

(1) a. B: Do you like Lubbock better than Dallas?
b. A: Yeah
c. B: Why?
d. A: Uh, because people are so much nicer

e. B: Right / Yeah / Okay / Really / No

The potential cue word responses in (1e) span meanings from strong agreement to disbelief. What speaker A does next will differ depending on whether B has accepted their assertion that people are nicer in Lubbock (e.g. with an affirmative) or whether B appears to be questioning it (e.g. with *really*). Impressionistically, the prosody of the response also seems to have an effect beyond the choice of cue word. This is reflected by the fact that individual cue words are usually associated with several different types of dialogue act in annotated corpora (Benus et al., 2007, Gravano et al., 2007). Being able to handle these types of utterances is crucial for wide coverage dialogue system development (Ward and Tsukahara, 2000, Edlund et al., 2009, Gravano, 2009, a.o.). As such, the majority of studies on the interpretation of cue words and their prosodic characteristics have been concerned with recognition of specific dialogue acts for the purposes of automated dialogue management. In particular, several studies have focused on disambiguation of affirmatives like *okay* and *yeah* as agreements or backchannels (Shriberg et al., 1998, Gravano, 2009, Truong and Heylen, 2010). This work is usually domain specific and there has not been much of an attempt to generalize the results or to link them to more general theories of prosodic meaning. So, we would like to see whether variation in cue word prosody and their discourse interpretation matches up with the predictions made by theoretical analyses of terminal rises.

While non-sentential utterances like cue words have not generally had a lot of attention from theoretical linguists, there have been many proposals about the contribution of prosody more in general terms. As discussed in Chapter 2, the contribution of intonation has often been framed in terms of discourse commitments and speaker attitude (Gunlogson, 2003, Haan, 2002, Steedman, 2000, a.o.). Similarly, attempts have been made to connect intonational form to the perception of bias in questions from the semantic point of view (Romero, 2006, Nilsenova, 2006, Reese, 2007). These sorts of studies visit similar ground to the empirical work on cue words in terms of attempting to identify what the speaker's intentions are for how the discourse will proceed. However, the more theoretical work tries to do this by modelling what specific discourse moves can be represented in a logical form, whereas the empirical work focuses on how we can identify those moves.

We would like to see whether these two approaches tackling the prosody-meaning map actually meet in the middle. Cue words make a particularly good test case for this because their contribution to discourse seems to be at the same sort of level as has been suggested for intonational components like terminal rises. As such, this chapter presents investigations into the prosody and interpretation of cue words in terms of dialogue acts in spontaneous conversational speech, with a particular focus

on the cue word *really*. *Really* seems to act at times a clarification type question requiring a response and at other times simply as a passive signal that the speaker is listening, i.e. a backchannel. So, a first pass hypothesis based on the theoretical work would be that rises associate with question type dialogue moves for this cue word.

This chapter is structured as follows. Section 3.2 introduces cue words, the dialogues acts associated with them and previous findings about their prosody. Section 3.3 investigates whether there are direct prosodic cues to the *really* backchannel/question distinction in the Switchboard corpus of conversational speech. While we find some general differences, they do not appear to be enough to separate the two categories in the prosodic feature space. Section 3.4 examines the contribution of contextual features to this disambiguation problem. Overall, it appears that terminal pitch rises do not cue a question interpretation. However, effortful prosodic features do increase the likelihood of the question label, as does evidence that the *really* in some way interrupted the discourse, e.g. an evaluative response, less speaker overlap. In Section 3.5 we discuss the implications of the findings from the corpus studies for theories of intonational meaning. Section 3.6 concludes.

3.2 Cue Word Responses

In this section, we will look at the distribution of cue word responses in the Switchboard corpus of spontaneous speech to get an idea of what they mean and what their discourse use is.[1]

3.2.1 Cue Words in the Switchboard Corpus

There is no absolute list of cue word responses. However, looking at the distribution of turns and dialogue acts in speech corpora, we can clearly see that some one word turns are much more frequent than others and have common properties with respect to dialogue acts. In the following we take advantage of several annotation efforts based on the Switchboard corpus (Godfrey et al., 1992). This corpus as a whole contain 2400 two sided telephone conversations of roughly 6 minutes in length, involving 543 speakers of American English of both genders, drawn from various parts of the

[1]Note: the cue word (cue phrase, discourse maker) label covers a broad class of linguistic expressions that are taken to be explicit indicators of discourse structure (Hirschberg and Litman, 1994). Cue words vary in what sort of role they play in a discourse. A broad distinction exists between cue words which act as connectives, e.g. *and*, *because*, *instead*, and those that act as responses, e.g. *right*, *okay*, *really*. In this chapter we will focus on the latter type, specifically one word responses whose use is primarily evaluative.

country. The goal of this collection was to elicit natural, spontaneous conversational speech. For each conversation, participants were given a topic to speak about along with some prompt questions, e.g. affirmative action, credit card use. The speech was fully transcribed and automatically time aligned at the word level. Hand corrected time alignment was later completed and released from ISIP.[2]

Subsets of Switchboard have been the focus of various annotations efforts. Several of these have been consolidated into a standardized XML format in the Switchboard in NXT project (Calhoun et al., 2010). This collection allows the corpus to be queried simultaneously over multiple layers of annotation. This includes tags augmenting the Discourse Annotation and Markup System of Labeling of Core and Allen (1997), SWBD-DAMSL (Jurafsky et al., 1997). This dialogue labeling scheme includes 42 tags clustered from the original 220 label set due to data sparsity (cf. Section 2.3.4 and Table 2.2). Tags were applied to previous hand segmented utterances (Meteer and Taylor, 1995). These basically attempt to model both the contribution of an utterance in terms of speech acts and also expectations about how one unit will be responded to by another (i.e. adjacency effects). Annotators worked on written transcriptions only.

Another set of dialogue act level annotations of Switchboard comes from the DARPA EARS simple metadata extraction project (SWBD-MDE Strassel, 2003). As for SWBD-DAMSL, the metadata produced in this annotation effort was produced with the goal of creating resources to support a development of automatic speech recognition and synthesis technologies. The metadata annotations were defined over SUs ('Semantic', 'Syntactic', or 'Sentence' units). The intention was to mark out units that expressed one thought or idea. Unlike the DAMSL based set, annotators were instructed to use both the transcript and the audio in identifying SUs and, moreover, to use syntactic/semantic information as the primary cue (as opposed to, for example, phrasing).[3] SWBD-MDE includes 4 types of SU: statement, question, backchannel, and incomplete. So, this is a much broader categorization than the DAMSL scheme. We look at how the differences in label sets are reflected for cue word responses in the next section. These differences in turn shed light on the different meanings and uses of these discourse markers.

3.2.2 Cue Words in SWBD-DAMSL

From the 642 SWBD-DAMSL annotated conversations accessible in Switchboard-NXT, 92% of syntactic one word utterances were labelled as syntactic interjections

[2] `http://www.isip.piconepress.com/projects/switchboard/`

[3] The ideal boundary placement is at the clause level. Annotators could also mark sentence internal clauses, e.g. if/then clauses, coordinations.

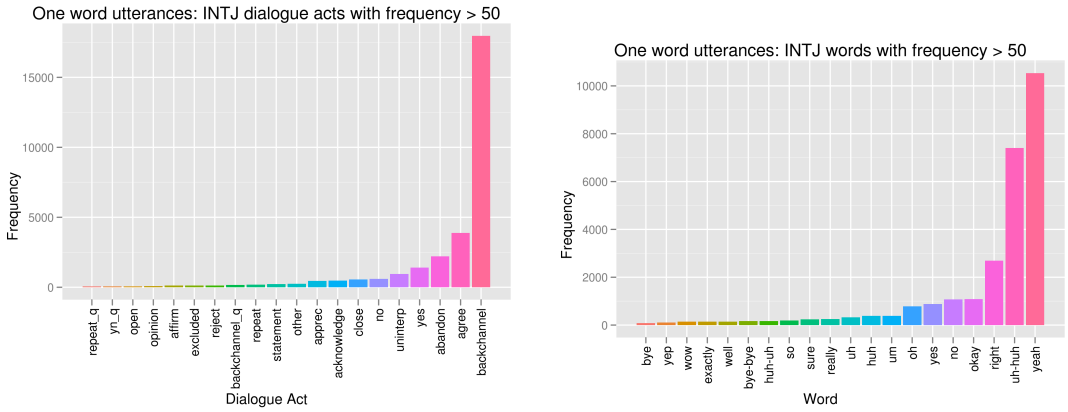


Figure 3.1: SWBD-DAMSL: One word turns.

(INTJ) at the top level. 77% of non INTJ utterances were abandoned turns, and 5% were statements (i.e. answers to questions). So, in the following we concentrate on INTJ as predominantly a discourse marker category. Figure 3.1 shows the frequencies of words and dialogue acts for one word interjections. We see that by far the most frequent cue word is *yeah*, followed by other affirmatives like *uh-huh*. In terms of dialogue acts, the common thread between these utterances appears to be their use as backchannels, simply signalling that the other participant should keep talking.

Figure 3.2 shows dialogue act frequencies for specific cue words. While all of these words have some backchannel usage, they vary in the frequencies of other uses. We can get a good idea about the range of meanings these cue words span by looking at the differences in these distributions. The affirmatives *yeah* and *right* are predominantly used as backchannels, but they are sometimes used as actual agreements. Although the distributions are similar, *right* is not used to answer questions (the YES act). Unlike, *right* and *yeah* one of the main uses of *okay* is as a response acknowledgement. This differs from the backchannel category in that they are acknowledgements of answers to questions.[4]

(2) A: How about um Pink Floyd? (sw2032)
B: Sure yeah i like them-
A: *Okay* so so then we got we got some of the same things because

A large number of *okay*s fall into the OTHER category. These seem to also be best characterized as general acknowledgements or acceptances of the state of the affairs. In the following, by saying *okay* speaker B seems to be acknowledging that its time

[4]The thing about DAMSL is that some categories encode local structure and some do not.

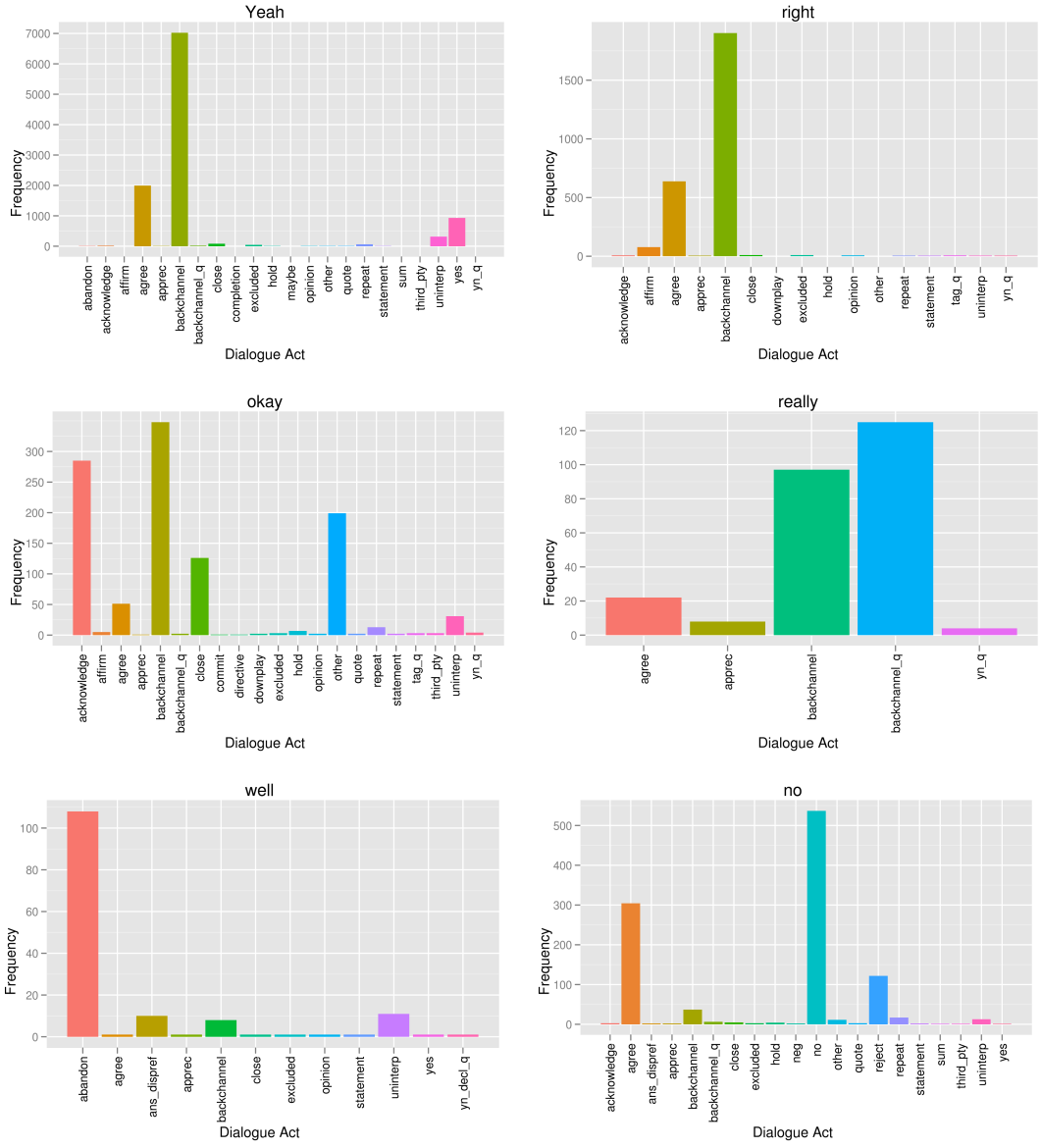


Figure 3.2: SWBD-DAMSL Dialogue acts for specific cue words.

to finish the conversation.

(3) A: I'm drawing the blank (sw2015)
B: [laughter] oh goodness okay is our five minutes up

Really is an example of a non-affirmative that is used as a backchannel. However, it's main categorization is as a backchannel question. In this annotation scheme this means it is a signal that the other speaker can keep speaking that happens to take the form of a question. The following gives examples of *really* labelled as an agreement (a rare annotation) and in the more frequent backchannel question category.

(4) B: I needed the money [laughter] this is five bucks here [laughter] (sw2039)
A: [laughter] I guess yeah I- I guess we all can
B: Yeah really$_{\text{agree}}$
B: Do you work for TI?
A: No no i work for GTE
B: Really$_{\text{bcq}}$ how'd you hear oh we're not even supposed to be talking about this though are we

The assumption here is that, although these utterances have question form, they do not actually function as questions – they merely signal that the speaker is attending to the conversation. The DAMSL manual[5] notes that they are 'marking these distinctly because we suspect that they will mess up the prosodic utterance detector if they are just thrown in with the b's, since they have question intonation'. However, Figure 3.2 shows that a substantial number of *really*s were labelled as backchannels rather than backchannel questions. It is not clear how this distinction was made or whether it is meaningful. Note that even though the annotators did not have access to the audio, a prosodic distinction may still be reflected in other ways, like turn overlap or the presence of other backchannels in the vicinity. This seems to be the case in the following example which is from the same conversation as the previous.

(5) A: Uh Turner Broadcasting (sw2039)
B: Uh-huh
A: Has been uh broadcasting all of the James Bond movies
B: Really$_{\text{bc}}$ uh-huh
A: Every night this week

Another non-affirmative cue word that has an interesting distribution is *well*. It

[5]http://www.stanford.edu/ jurafsky/manual.august1.html

was most frequently marked as an abandoned move. This suggests that these had something of a discourse connective use. Like *really*, *well* seems to express an inability or unwillingness to admit the utterance at issue into the common ground (Schiffrin, 1994). So, it seems that speakers use *well* when they want to signal some complication in the dialogue. This fits with the sometime marking of the cue word as indicating a dispreferred answer.

(6) A: and i don't believe any of us would have to purchase any extra vacation days if they did that (sw2062)
B: $\text{well}_{\text{abandon}}$ true true
A: uh

Finally, we see that, as we would expect *no* usually means NO, i.e. a negative answer to a question, but it is also used as an agreement, e.g. for negative statements. Overall, we see that cue word responses have a wide range of uses in dialogue. Moreover, they seem to apply to variable levels of grounding (Allwood et al., 1992, Clark, 1996). The common backchannel use appears to act at the lowest level, simply signalling attention or contact. However, other uses act at the highest level in that they give the speakers actual evaluation of the utterance they are responding to.

3.2.3 Cue Words in the MDE Annotations

We can compare the DAMSL DAs to the metadata extraction annotations (MDE) from the Switchboard component of the English CTS Treebank with Structural Metadata Corpus (Strassel, 2003, LDC2009T01). As in the SWBD-DAMSL set, the common thread between the high frequency one word SUs was their use as backchannels. In this annotation effort this backchannels were defined as 'words or phrases that provide feedback to the dominant speaker by indicating that the non-dominant speaker is still engaged in the conversation (though not actively participating at the moment)'. So the main distinction between backchannels and the other categories is more or less whether the semantic content of the SU can be ignored or not.

Figure 3.3 gives the raw counts for one word SUs in the MDE annotation. We see the same pattern as we did for the one word turns in the DAMSL annotation, with *yeah* as the most common word. Figure 3.4 shows the distribution of SU types for the cue words we looked at in the previous section. Again, the pattern is basically the same as for the DAMSL set, but we lose granularity of interpretation. The vast majority of affirmatives *yeah*, *right* and *okay* were labelled as backchannels. On the other hand, *no* usually gets interpreted as having an active negative contribution. However, there is no clear distinction between backchannels, agreements, and answers.

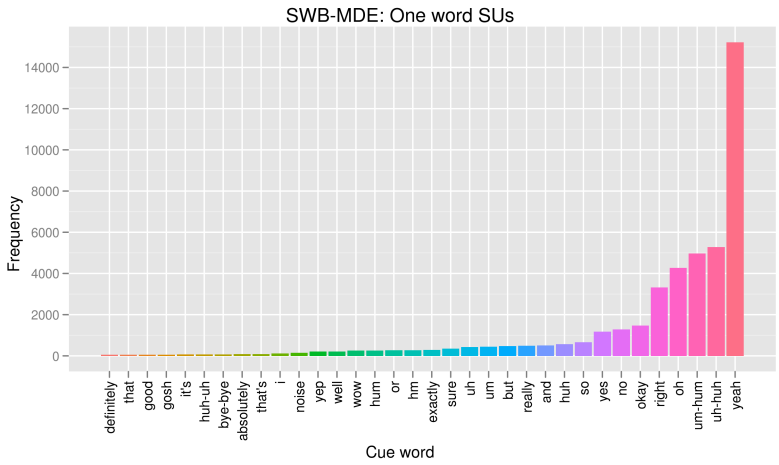


Figure 3.3: One word SUs in Switchboard MDE with more than 50 observations.

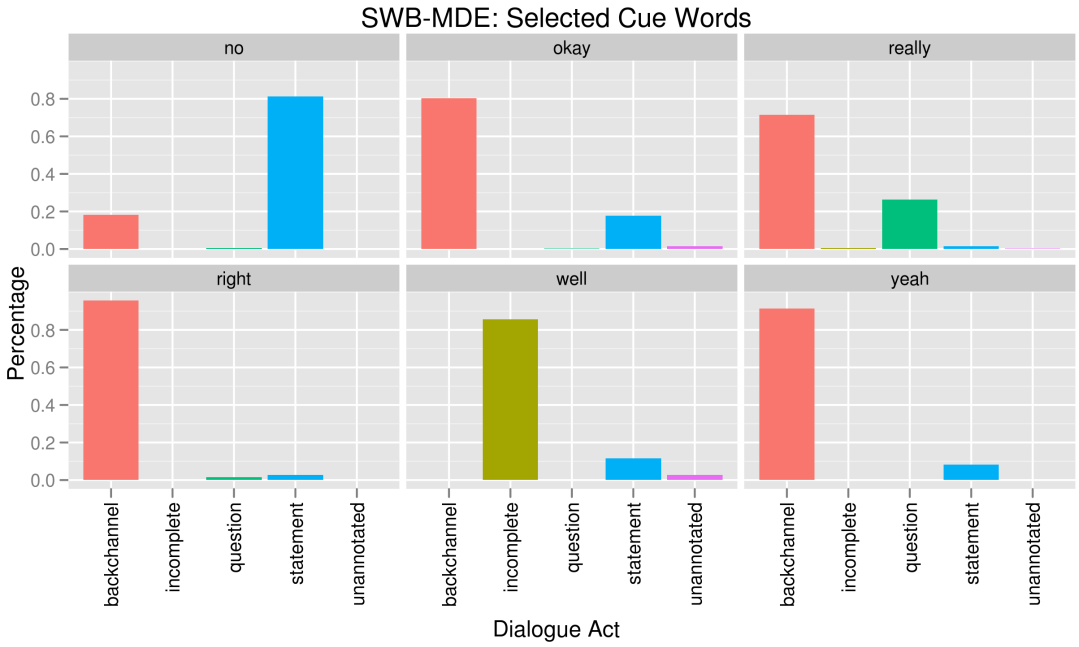


Figure 3.4: SWBD-MDE SU types.

The questioning side of *really* again contrasts to the backchannel interpretation. It appears that cue word *well* was treated as a connective: when it appears alone, it seems incomplete. From this we can glean that the discourse contexts where non-affirmatives can be treated as passive contributions are less frequent that those for affirmatives.[6]

Looking at the dialogue acts associated with cue word responses, it appears these utterances act at the discourse level. They guide what enters the common ground and speakers' public beliefs. Backchannels appear to be a sort of bleached version of the cue word (more like a grunt). That is, an utterance that accepts the contents of an utterance without expressing strong agreement or disagreement. However, backchannels of different types are not necessarily interchangeable. Changing the response in example (7) expresses different evaluations of B's previous utterance, even though A's actual response *oh really* response was marked as a backchannel. So, it appears that backchannels convey more information than a mere acknowledgement that an interlocutor has the floor.

(7) B: so uh the uh the Plano area and the Richardson area probably are are about the same
A: oh really / right / uh-huh (sw2017)

Neither of the annotation schemes we have looked at make predictions about the relationship between prosody and dialogue acts, although the DAMSL guidelines seem to assume that some categories that will have distinct prosodic features by default (i.e. backchannel questions). In the MDE annotation FAQ, an *oh really* example is given to show that prosody does not always disambiguate these types and the default should be to go with the backchannel. This is goes somewhat against the rest of the annotation guidelines which actually seem to bias away from marking backchannels.

3.2.4 Previous Work on Cue Words and Prosody

Several studies have investigated the prosody of cue word type responses in order to improve backchannel detection and turn-taking in automated dialogue systems, the

[6]Note: The MDE guidelines make a distinction between define backchannels and discourse markers, where the latter are 'inherently active contributions to the discourse, and signal such activities as a change in the speaker, taking or holding control of the floor, giving up the floor or the beginning of a new topic.'. Examples of discourse markers under this scheme include cue words like *yeah* and *okay*. The guidelines state that backchannel words that preface a longer statement should be marked as discourse markers rather than as backchannels. However, it is not clear whether this was actually done. There only appear to be 5 *yeah* DMs in the corpus, while there are 249 *yeah* initial SUs with more than one word.

former of which is usually cast as a dialogue act disambiguation task. Most of the work in this area has focused on affirmatives, where the main distinction is between the passive backchannel sense and a stronger agreement signal. Correlations between prosody and DA types have been found in individual studies, however results seem to vary with the dialogue type.

There are basically two types of results: those that associate backchannels with 'quiet' features, and those that link them to more marked features. The latter type of result come from from studies based on more task-oriented data. In particular, detailed studies carried out on affirmative cue words in the Columbia Games Corpus found that higher mean pitch and intensity and greater pitch slope as distinguishing features of backchannels (Benus et al., 2007, Gravano, 2009). Looking specifically at the uses of *okay*, Gravano et al. (2007) found pitch rises to be the only strong correlation to backchannel perception when factoring in turn latency features, suggesting that the rise is in fact a strong cue to this move type. The correlation between rises and backchannels has also been found in other tasks oriented dialogues in English (Hockey, 1993), and cross-linguistically, e.g. map task dialogues in Bari Italian (Savino, 2011),[7] Swedish (Heldner et al., 2008), and Dutch (Caspers, 2000).

The quieter characterization of backchannels come from more conversational data, for which there have been far fewer quantitative studies. Nevertheless, Shriberg et al. (1998) found that backchannels had shorter duration and lower mean intensity than agreements in the SWBD-DAMSL. Truong and Heylen (2010) found a similar distinction between backchannels and assessments (i.e. agreements) looking specifically at the use of *yeah* in the AMI meeting corpus. Prasad and Bali (2010) also characterizes backchannel uses of the Hindi affirmative 'hã' *yeah* in telephone conversations as having shorter duration, low power and relatively flat pitch contours compared to agreements. Similarly, Ward (2004) describes backchannels as being quiet and having flat pitch in a qualitative study of conversational feedback from varied settings.

These differences suggests a style/genre divide in how prosody is interpreted. This is something that doesn't mesh easily with theories of prosodic meaning which generally assumed to be more robust to situational issues. Although there is more than a little fuzziness in the definition of a backchannel, the common thread is that they only really signal that the speaker is attending the conversation. For example, Gravano (2009) describes backchannels in the Columbia Games Corpus as indicating 'only "I'm still here / I hear you and please continue"'. So, accounts where rises signal contingency (Gunlogson, 2008), a propositional test (Nilsenova, 2006), or more generally speaker responsibility (Steedman, 2000) would seem to exclude a backchannel interpretation since they produce conditions in the discourse beyond just allowing the

[7]Note, they take backchannels to be uses that don't precede a speaker stay, which is more of a continuer definition than a passive one.

hearer to keep talking (cf. Section 2.4). Similarly, cases of backchannels bearing rises seem to be troublesome for theories that associate rises with extra attitudinal type meaning, e.g. uncertainty. We might expect that any extra contribution would push the cue word out of the backchannel category.

The difficulty with looking at the connection between affirmatives, dialogue acts and rises is that it is not clear what category the affirmative would fall into if the rise did add an adverbial type meaning. Moreover, the overall difference in meaning between an agreement, an acknowledgement and a backchannel is rather subtle. A continuation or a change in topic after an agreement is just as felicitous as after a backchannel, within the bounds of relevance. So, operationally it is hard to see the effect that a rise would have. A potential source of more direct differences in interpretation come from non-affirmatives. Our candidate here is cue word *really* whose DA meaning spans the backchannel/question range. The difference between these two acts is much more stark. Unlike backchannels, questions inherently make an active contribution to the shared discourse structures: they set the QUD. Similarly, the attitudinal predictions seem more inline with this dimension. The more the speaker sounds uncertain, the more the *really* should seem like a real clarification request. In operational terms, we would expect that question *really* should act as a clarification question requiring explicit resolution. So, changing the topic would incur a greater penalty in this case. In contrast, the hearer should be able to move on after backchannel *really* and not seem uncooperative. We take a more detailed look at how the prosody of this cue word in terms of with dialogue act annotations and theoretical predictions in the next section.

3.3 Corpus Study: *Really*, Backchannels and Questions

3.3.1 Goals and Hypotheses

This section presents a corpus study of cue word *really* in conversational speech. Recall that one of our overarching goals was to find appropriate response variables for prosodic variation on cue words. The following studies investigate how well dialogue acts do for this purpose. More specifically, we want to see whether we can associate specific prosodic features with specific dialogue acts. As a first step we focus on the relationship between the DA categorization and the prosodic features of the target *really*s in isolation. We will return to the contribution of contextual features in Section 3.4. The major distinction we saw previously for *really* is between the backchannel and question categories. We can frame this problem as follows:

(8) How do we know when to treat *really* like a question?

The taxonomies of non-sentential utterance generally assume fairly sharp distinctions between question and assertion/acknowledgement type categories (Fernandez, 2006, Schlangen, 2004). So this work bears on the question of whether we can indeed make clear distinctions between dialogue act types, especially when the contrast involves a low information category like backchannel. If prosodic features really make the distinction between categories, we would expect a sharp contrast in the feature distributions conditional on those categories. Contingency (Gunlogson, 2008), epistemic test (Nilsenova, 2006), cognitive uncertainty (Reese, 2007), and hearer responsibility (Steedman, 2000) analyses of rises predict a strong link between rises and questionhood, since these in these analyses rises add an explicit post-condition to the discourse (cf. Section 2.4).[8] For example, the Nilsenova's modal analysis claims that rises acts as a test on common ground. The predictions are less clear for a non-finality type analyses (Bartels, 1999), since both backchannels and questions suggest discourse openness on some level.

We also need to consider the contribution of the pitch accents attached to rises/falls. These are also assumed to contribute discourse level meaning in analyses such as Pierrehumbert and Hirschberg (1990), Steedman (2000) and Büring (2003). In fact, Hirschberg and Litman (1994) use this to explain why cue phrases such as *now* tended to exhibit low pitch in their discourse use: it is assumed that L* marked items do not add semantic content to participants beliefs. However, the predictions are not totally clear. When we apply this to the non-affirmative *really* responding to proposition p, we predict that the falling accent (H*L%) version should express more bias towards really(p) being added to the common ground which would suggest a backchannel interpretation. For the rise (L*H%), however, it could go either way. L*H% might mark the cue word as not adding anything (i.e. a backchannel) or it might lead to a questioning interpretation. The latter is the prediction of Steedman's account where H* and L* accents mark informational units that are agreed upon and contentious respectively.

We expect the backchannels to exhibit lower intensity and generally less effortful pronunciation based on the previous studies of conversational speech, particular around Switchboard. This is in line with the Effort Code (Gussenhoven, 2002).[9] Based on the literature then, our two main hypotheses are as follows:

(9) a. Lower mean intensity, lower mean pitch, shorter duration $\mapsto$ Backchannel.

[8] Recall too that Gunlogson (2003) and Nilsenova (2006) only deal with nuclear rises.

[9] Note, however, that both categories are compatible with the Frequency Code: submissiveness linked to high boundary pitch could be interpreted as ceding evaluation of a proposition (i.e. the question interpretation) or simply ceding the floor (i.e. a backchannel).

MDE SU	-	oh	other	Total
backchannel	210	227	29	466
question	84	67	0	151

Table 3.1: SWBD-MDE: *Really* SUs, by preceding cue word.

b. Terminal rises, increased convexity $\mapsto$ Question (MDE), Backchannel Question (DAMSL)

If we find, instead, that backchannels are as equally rising as question, this would be strong evidence against accounts which claim that rises act on the proffered content of their carrier utterance (contingency, test, hearer responsibility). If we find a significant number of high falling *really*s as questions, this would be evidence against and H*/L* distinction in terms of predication, or whether the informational unit is agreed upon. A result showing the same prosodic characterization of backchannels as reported for task-oriented dialogues, where backchannels are rising but questions are not, would be very difficult to explain and we would have to consider strong lexical effects to be at play. If the expected result prevails, we have some more evidence of a higher level genre effect.

3.3.2 Data and Method

In order to map prosodic features to dialogue acts, we need to choose a set of acts to deal with. In the following study, we will look at *really* in the Switchboard corpus from the point of view of both the MDE and DAMSL labelling schemes. The MDE guidelines focus on function rather than form. The defining quality of a backchannel in this scheme is its passiveness. In contrast, the DAMSL scheme has a mix of form and function criteria in its guidelines, and the major the major distinction for *really* is between the backchannel and the backchannel question categories. The MDE style of annotation is of more relevance to our goals. However, looking at the DAMSL categorization should give us an idea of what it means to have question form and what sort of consequences this has for the discourse.

617 backchannel and question *really* SUs were extracted based on the SWBD-MDE 2003/04 annotations described above. 472 of these had only one word in the SU. 181 were preceded by another backchannel: 149 *oh*s, 25 *yeah*s, 3 *really*s, and 1 *I mean*. 145 *oh really* (two word) SUs were also identified and added to the data set. This resulted in 466 backchannels and 151 question *really* tokens (see Table 3.1).

We also examined the prosodic characteristics of 495 *really* interjections from the SWBD-DAMSL annotations (via Switchboard in NXT (Calhoun et al., 2010)). As in

DAMSL Dialogue Act	-	no	oh	really	uh-huh	yeah	Total
acknowledge			1				1
agree	9			2	1	10	22
apprec	8		3				11
backchannel	81		72	3	1	6	163
backchannel_q	120	3	166		1		290
close			1				1
excluded			2				2
yn_decl_q			1				1
yn_q	4						4

Table 3.2: SWBD-DAMSL: *Really*s Dialogue Acts, by preceding cue word

the MDE set, these *really*s were sometimes preceded by another cue word as shows in Table 3.2. The majority of the *really*s were labelled as backchannel questions, but about a third of the tokens were labelled as plain backchannels.

Features

The word boundaries were taken from the corrected ISIP transcripts used in the MDE annotation files and the Switchboard NXT release for the SWBD-DAMSL portion. F0 and intensity features were extracted via Praat for each of the identified *really*s. For the F0 data, input parameter range values for Praat were set based on estimated pitch range for each token (Evanini and Lai, 2010). The extracted values were then converted to semitones based on each speaker's F0 median for the conversation. The contours for each token were smoothed using a Butterworth filter with a normalized cut off frequency of 0.1. Intensity data was normalized by speaker to z-scores. The duration of each word was also normalized to z-scores based on speaker mean and standard deviation for 2 syllable words in that conversation.

The following aggregate statistics were calculated over each *really*: mean, standard deviation, linear regression slope, range, maximum, minimum, jitter, relative times of the maximum and minimum (proportionally). We also fit order 4 Legendre polynomials to the pitch and intensity contours, giving 5 coefficients (Kochanski et al., 2005). We are particularly interested in the first three coefficients which indicate the overall height, tilt and convexity of the polynomial. The first five Legendre polynomials are shown in Figure 3.5.

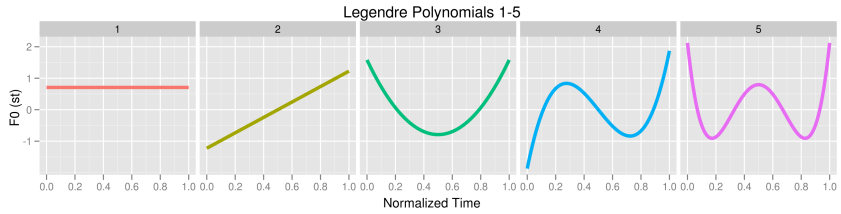


Figure 3.5: Legendre Polynomials

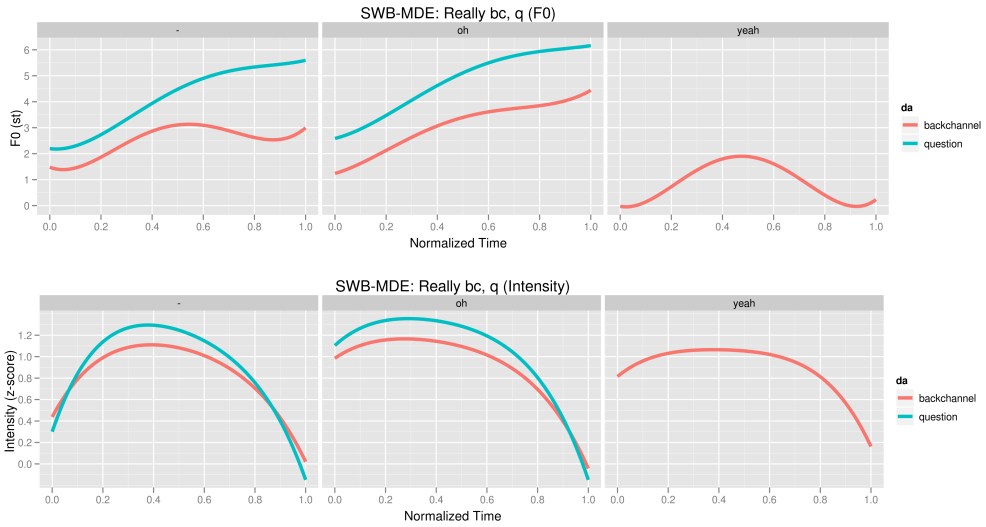


Figure 3.6: Means contours, by immediately previous cue word.

3.3.3 Distributions in SWBD-MDE

Data Exploration

Contours generated from the mean values of the Legendre coefficients for backchannels and questions are shown in Figure 3.6. Looking at the data this way it seems that the main difference between backchannels and questions is one of height rather than of whether or not there is a terminal rise. Similarly, questions on average appear to have greater intensity peak around the first syllable. So, it seems that *really*$_q$ is produced with greater effort. Interestingly, we see that *really*s prefaced by a *yeah* are generally situated lower than the more frequent singleton and *oh really*s. All of these tokens were labelled as backchannels, which makes sense compositionally: the *yeah* already signals acceptance.

We look at the effect of the various prosodic features on dialogue act labelling by

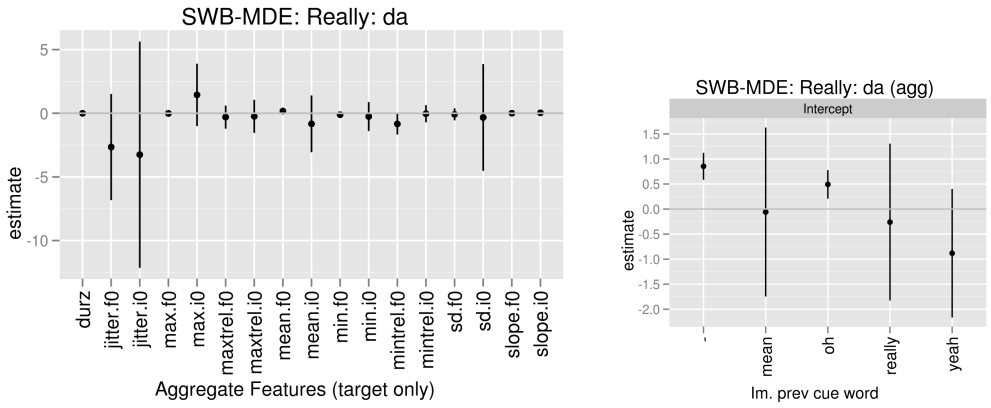


Figure 3.7: Parameter estimates for aggregate prosodic features and the immediate previous cue word (± two standard errors).

fitting a multilevel logistic regression model of the following form (Gelman and Hill, 2007):

(10) $$\Pr(\text{da}_i = \text{question}) = \text{logit}^{-1}(\beta^0 + \text{prosfeat}_i + \alpha^{\text{prev.cw}}_{j[i]}), \text{ where}$$

$$\alpha^{\text{prev.cw}}_j \sim N(0, \sigma^2_{\text{prev.cw}}) \text{ for } j = 1, \ldots, 5.$$

The dependent variable is probability that observation i gets tagged with the question label (i.e. backchannel=0, question=1). The group level parameter $\alpha^{\text{prev.cw}}_j$ represents the effect of having an immediately previous cue word like *oh* or *yeah*. In these models, prosfeat_i represents a set of prosodic features (i.e. individual level predictors) and coefficients associated with token i. Model parameters were estimated using R package `lmer`. Figure 3.7 shows parameter estimates for the model based on aggregate features. None of the estimates for the prosodic features included in the aggregate model were more than two standard errors from zero (our bar for significance). We do see significant effects for the immediate previous cue word. However, this basically represents the fact that the small number of *really*s prefixed by a cue word other than *oh* are all backchannels. When we take into account the intercept estimate (-1.96, s.e.=0.91), we find that this model basically predicts all tokens in the data set to be backchannels.

Figure 3.8 shows the parameter estimates for the model based on the F0 and intensity Legendre coefficients and z-scored duration. The essential difference between this model and the one based on normal aggregate features is the presence of the

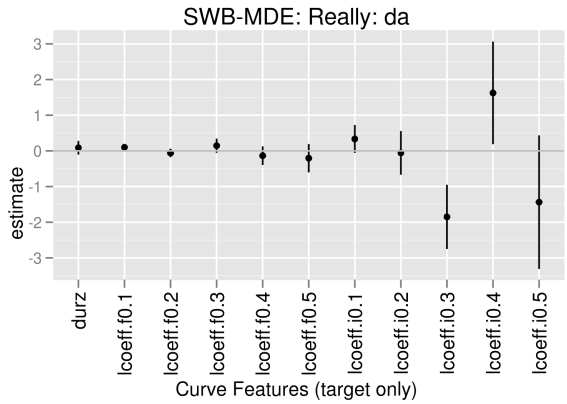


Figure 3.8: Parameter estimates for curve features.

curvature features, i.e. the third Legendre coefficient. The estimates suggest that these features are in fact useful for this task. For this model, we don't see any significant effects for the immediate previous cue word. However, the estimates for the first F0 Legendre coefficient, and third and fourth intensity coefficients were significantly different from zero. This reflects what we saw for the mean contours shown in Figure 3.6: *really*s with higher overall F0 and more concave intensity contours were more likely questions. Similarly, the positive estimates for lcoeff.i0.4 indicates a somewhat raised intensity at the contour end for questions. The fact that the estimates for the intensity parameters are so much bigger than F0 parameter is due to differences in scaling. If we take the contribution from the median lcoeff.f0.1 height value we get $0.10 \times 3.4 = 0.34$, for lcoeff.i0.3 we get $-1.85 \times -0.29 = 0.54$. Overall, we don't really see any effect of terminal rise in determining whether a *really* gets a backchannel or a question interpretation. Instead the data support the idea that question *really*s are produced with more effortful prosody in general. However, the estimated effects are quite small.

How does the data vary?

Figure 3.9 shows the distribution of Legendre coefficients with respect to lcoeff.f0.2 (tilt), lcoeff.f0.3 (convexity). We also show the mean contours (from Legendre coefficients) for the data points in each quadrant of the plane. It appears that the *really*s basically vary continuously through a range of contour shapes: concave falls (lower left quadrant), convex rises (upper right), rise-plateaux (upper left) and flat contours (at the origin). While the questions do tend to inhabit regions representing the higher, rising contours in these planes, there is a good deal of overlap in these

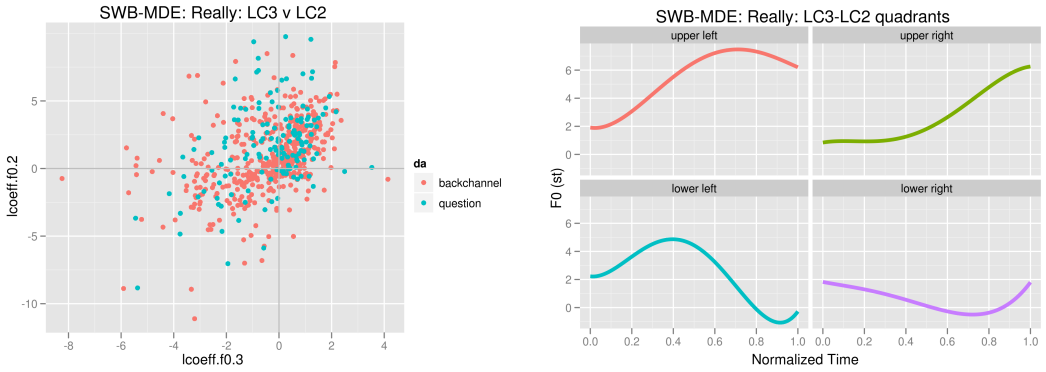


Figure 3.9: F0: convexity (lcoeff.f0.3) v tilt (lcoeff.f0.2), mean contours for each quadrant (right).

(a) F0 Height v Tilt

(b) PCA: F0, Intensity, Duration

Figure 3.10: MDE *Really*: Height, Tilt, and aggregate feature PCA projection.

areas with the backchannel observations.

Principal Components Analysis (PCA) of the curve features (F0 and Intensity Legendre Coefficients and normalized duration) indicated that the data varied mostly in terms of F0 tilt and convexity, and next in terms of intensity convexity and F0 height. Figure 3.10 shows the data projected onto the first two principal components. Once again, the presence of F0 tilt or convexity features (x-axis) does not seem to be involved in the DA distinction. However, we do see a somewhat clearer distinction between DA distributions on the y-axis (compared to the LC1 v LC2 graph). More negative values in PCA graph basically indicate higher F0 height and intensity, which were the features picked out by the previous logistic regression. However, the overlap between groups is still quite pervasive. Overall, it seems that while questions do sit

da, da.pred	backchannel	question
backchannel	453	11
question	127	24

Table 3.3: Predicted DA from the curve feature logistic regression

at the more effortful end of the prosodic feature scale, there is a lot of variation in the production of *really* that isn't reflected in the dialogue act/SU annotations. We will return to what this variations actually means for interpretation later. First, however, we would like to see whether the features associated with *really* questions, or other features, can actually be used to separate the dialogue acts.

3.3.4 Prosodic Separability n SWBD-MDE

The logistic regression model based on the curve features correctly predicts more questions than the model based on aggregate features (which predicted none). However, the accuracy is very low for the question class (Table 3.3). So, although there are general differences between the backchannels and question groups, it is not clear whether these are enough to be able to distinguish one class from another. This section presents machine learning experiments with the goal of determining whether the dialogue act classes are separable based on prosodic features, and if so which prosodic features do the job.

Two types of classifiers were used to test this hypothesis: Decision Trees (DT) (`j48` in `RWeka`) and Support Vector Machines (SVM) classifier with radial basis function kernel (`libsvm` via R), with parameters determined by grid search. Decision trees have been used in previous work on dialogue act detection such as Shriberg et al. (1998) because they provide posterior probabilities on classification which can be incorporated into larger models. Since we can look at the decision tree structure directly, they can also give an indication of which features are important to the classification process. We compare the decision tree results to SVMs, treating the later as a sort of upper bound on the classification ability.

Table 3.4 shows the median cross-validation error rate and F-measure for classifiers based on the aggregate and curve features sets. Medians and 95% confidence intervals were calculated over 100 different randomizations of the data. The SVMs for the aggregate feature set appeared to produced the majority class baseline classifier, putting all tokens into the backchannel category for every iteration. None of the other classifiers performed significantly differently from that baseline. We can note that the DT classifier derived from the whole data set also resulted in the trivial decision rule.

	Error	95 CI	Fmeasure	95 CI
Base	24.55		0.649	
DTree agg	26.21	(24.72, 28.29)	0.655	(0.640, 0.670)
DTree lc	26.32	(24.63, 28.30)	0.669	(0.651, 0.689)
SVM agg	24.55		0.649	
SVM lc	24.71	(24.47, 24.96)	0.648	(0.647, 0.652)

Table 3.4: SWBD-MDE cross-validation results using all the data (464/151).

	Error	95 CI	Fmeasure	95 CI
Base	50.00		0.331	
DTree agg	43.58	(38.38, 47.74)	0.561	(0.516, 0.614)
DTree lc	42.17	(37.62, 47.30)	0.576	(0.534, 0.623)
SVM agg	40.85	(35.76, 46.35)	0.591	(0.527, 0.643)
SVM lc	38.69	(34.93, 43.06)	0.612	(0.550, 0.612)

Table 3.5: Cross-Validation results downsampling the data (151/151)

So, it seems that the two DA categories are not separable based on prosodic these features.

The mean differences we saw previously arise from the fact that the question class occupies a subspace of the feature space of the backchannel productions. Question *really*s aren't usually low in F0 or intensity, but backchannels can and often are high. We can see this if we train classifiers where the backchannel set has been randomly downsampled to match the size of the question set (100 different samplings) Table 3.5 shows gives median cross-validation results which shows that in this case classifiers do better than the baseline. This is because some downsampled sets exclude more of the backchannel tokens that are featurally similar to the question reallys. This sort of difference is evident in Figure 3.11 which shows F0 height (lcoeff.f0.1) and intensity convexity (lcoeff.i0.3) for the downsampled set with the best median error rate ($\approx$ 35% error rate) compared to that of the full set. So, if a *really* utterance has more effortful features, it is more likely to be given as a question label, but these features don't separate out question and backchannel SU classes. In terms of terminal rises, we see that not only can a *really* be interpreted as a question without one, but it can be interpreted as a non-question with one.

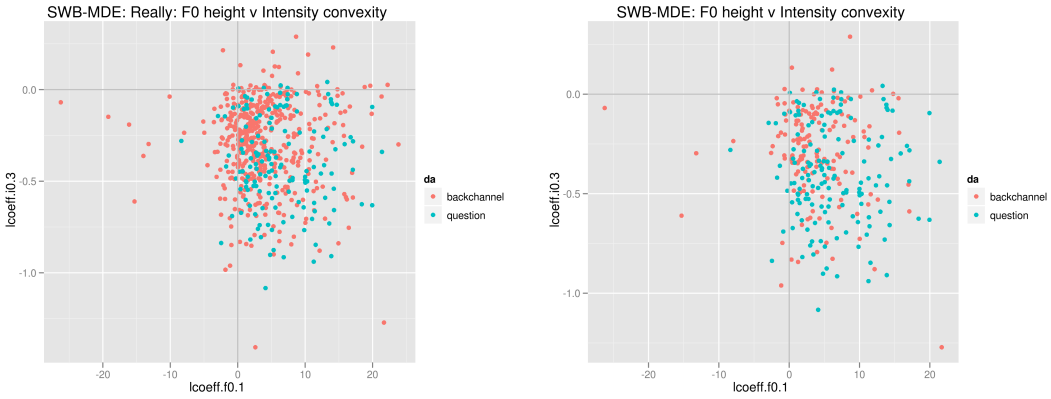


Figure 3.11: F0 height vs Intensity Convexity for (left) the whole data set and (right) a 'good' downsampled set.

3.3.5 Distribution and Separability in SWBD-DAMSL

It seems that differences the MDE SU labels don't reflect prosodic difference on *really*. However, the lack of a match up may simply be due to the fact that the categorization is just too broad. This suggests that we might fare better with the more detailed SWBD-DAMSL label set. In this annotation scheme cue word *really* was labelled as a backchannel and a backchannel questions. Items in the latter category were described as backchannels that had question form, which suggests a prosodic difference between categories. However, the annotators didn't have access to audio, so it is unclear how this worked in reality.

Prosodic features were extracted in the same manner as described in Section 3.3.2. Figure 3.12 shows tilt and convexity coefficients for the DAMSL dialogue acts associated with *really* interjections (cf. Table 3.2). The data has basically the same overall distribution as the MDE *really*s with most utterances having positive tilt. The backchannel question category seems to extend further into the rising space, although again there is more or less complete overlap between the backchannel question and backchannel categories. Figure 3.13 shows time normalized contours for backchannel, backchannel question and yes-no question *really*s preceded by *oh*, *yeah* and no cue word. As with the MDE *really*s we see that the backchannel questions have a more peaked intensity contour. We don't see the same degree of separation with respect to F0 height. However, we see more of a difference here between the DAs between the singleton *really*s, but not for the *oh really*s in terms of F0 tilt.

Figure 3.14 shows parameter estimates for multilevel logistic regression (0=backchannel, 1=backchannel question). In this case, we allow the effect of F0 tilt to vary with the immediately previous cue word. None of the estimates associ-

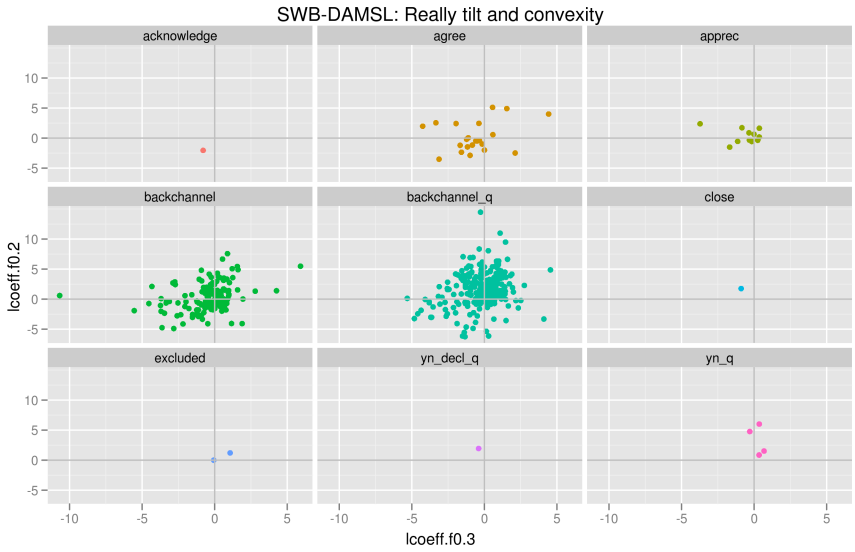


Figure 3.12: *Really* in SWBD-DAMSL: F0 tilt (LC2) and convexity (LC3)

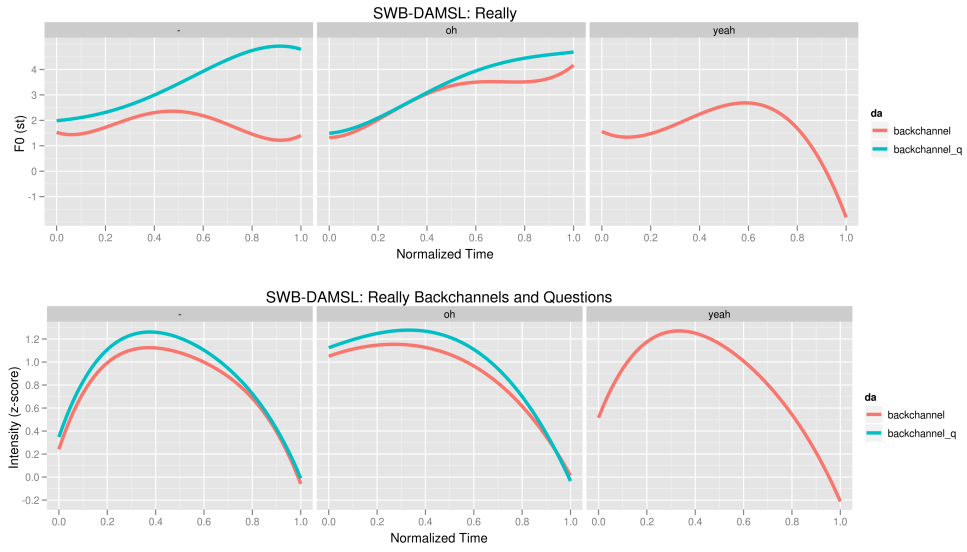


Figure 3.13: *Really* in SWBD-DAMSL: mean F0 and intensity contours

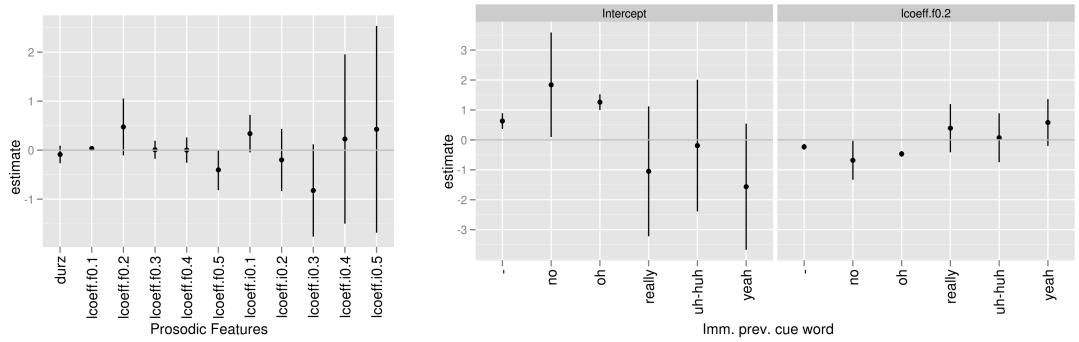


Figure 3.14: Parameter estimates for logistic regression (backchannel vs backchannel_q)

	Error	95 CI	Fmeasure	95 CI
Base	35.98		-	
DTree agg	38.31	(35.97, 41.19)	0.580	(0.550, 0.610)
DTree lc	38.80	(36.42, 41.28)	0.560	(0.573, 0.625)
SVM agg	36.07	(35.20, 37.09)	0.514	(0.503, 0.527)
SVM lc	37.04	(34.86, 39.41)	0.600	(0.573, 0.624)

Table 3.6: DAMSL: All data

ated with the prosodic features (as individual level effects) were significantly different from zero, although the estimates are in the direction we would expect: higher F0 tilt, and higher average intensity increase the probability of a backchannel question label. When we look at the effect of immediately previous cue word, we see that *really*s and *oh really*s are more likely to be classified as backchannel questions than backchannels, however, having positive F0 tilt doesn't actually increase probability of being labelled as a backchannel question. So, once again we don't get any clear prosodic distinction between the DAs with respect to pitch features.

Machine learning experiments give similar results to what we saw for the MDE data set. Cross-validation error rate is no better than the baseline for decision tree and SVM classifiers trained on aggregate or polynomial decomposition features, although we do get better results when we downsample the classes to equal sizes (Tables 3.6 and 3.7). So, under the DAMSL scheme, *really* backchannels seem to form a prosodic subset of the *really*s labelled as backchannel questions.

	Error	95 CI	Fmeasure	95 CI
Base	50.00		0.331	
DTree agg	41.46	(37.31, 45.78)	0.583	(0.538, 0.626)
DTree lc	39.12	(35.38, 44.00)	0.603	(0.555, 0.642)
SVM agg	39.73	(37.12, 42.49)	0.601	(0.548, 0.659)
SVM lc	40.72	(36.81, 45.40)	0.593	(0.537, 0.639)

Table 3.7: DAMSL: Cross-validation results for downsampled data sets.

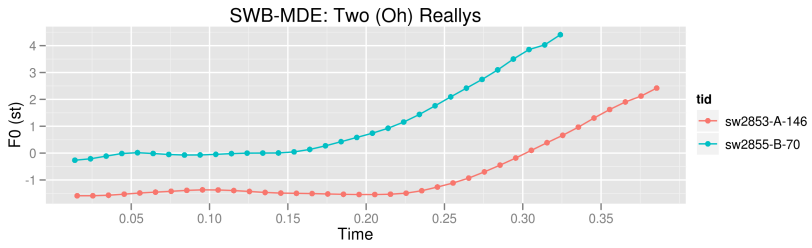


Figure 3.15: Two *really*s: sw2853-A-146 was labelled as a backchannel, sw2855-B-70 as a question in SWBD-MDE.

3.3.6 Summary

So far it does not seem that prosodic features, specifically terminal rises, make a sharp contribution to interpretation as reflected in dialogue act labelling. From the classification experiments, we don't see any clear separation of prosodic features between backchannels and questions for the SWBD-MDE *really*s, or for backchannel and backchannel question classes for the SWBD-DAMSL *really*s. While the questions in the SWBD-MDE set were in the more effortful area of the prosodic feature space, this turned out to be a subspace of the features space spanned by the backchannels. For the SWBD-DAMSL set, the less populous backchannel category overlapped with the less effortful end of the backchannel question set. So, at this point it seems that the prosody of the target word only has a very weak effect in determining what dialogue act label it receives. In particular, we don't see any connection between rising features and the SWBD-MDE question label. This is illustrated in Figure 3.15 which shows two tokens from the SWBD-MDE set with different labels but very similar tilt.

We do see something of a link between the F0 tilt and backchannel questions for the SWBD-DAMSL set, but this seems to be specific to the singleton *really*s. We can at least conclude that rises alone do not have any across the board effect on discourse

interpretation. This is not good news for analyses that give prosodic forms specific meanings that compose at the semantic level. If we assume that the underlying contribution of a terminal rise is to signal, for example, uncertainty, why would we get such different dialogue act interpretations as backchannels and questions in such similar prosodic space (cf. Figure 3.15)? We also saw that *really*s with falling accents could also get the question label, which doesn't quite work with the idea that H* accents mark agreed upon informational units.

However, effortful features do seem to increase the likelihood that a *really* will be identified as a question, in form or function, even if they don't necessarily turn an utterance into one. So, rather than treating rises like operator which performs an action on its carrier, it seems that we need to look at the contribution of prosody to dialogue act disambiguation more probabilistically. Moreover, we would like to know how the effects of the prosodic features of the target word stack up against effects of features drawn from the context. We take this up in the next section.

3.4 Corpus Study: *Really*, DAs and Contextual Features

Having looked at the relationship between the prosody of the cue word *really* and its interpretation at the dialogue act level, we now extend the analysis to examine the relationship between these categories and local contextual features. In particular, we want to see what effect the cue word has on what is to follow.[10] If the key characteristic of a backchannel is its passivity, we would expect that getting a response would tip the scales towards a question interpretation. Similarly, we would expect that large negative latency between the end of the turn which starts previous to the cue word, i.e. overlap, would indicate that the *really* is not being interpreted as questioning. The examination of the prosodic features of target utterance suggested that tokens with more effortful features were more likely to get the question label. In these cases, features like F0 height were relative to speaker means. However, we also want to consider effort relative to the other speaker. Intuitively, a *really* in a higher F0 key may not seem so effortful if the other speaker is also speaking in a high key.

Similarly, it is possible that the listeners might lean towards a backchannel interpretation if the previous speaker signals that they are expecting one. However, other studies have suggested that what may constitute a feedback eliciting signal might vary across dialogue types. For example, Gravano (2009) found a tendency

[10]Remember that while conversational participants only have the previous speech to work with, annotators have also have access to what comes next.

Set	Description
nsu	First SU starting after the target.
nwsu	SU of the first word starting after the target.
psu	Last SU ending before the target by the other speaker.
pwsu	SU of last word starting before the target by the other speaker.
Feature	
da	Dialogue Act
ts (stay, switch)	Speaker switch
evals (False, True)	Whether the next word is an evaluative by the hearer: yeah, yes, uh-huh, um-hum, okay, no, yep.
nw	Number of words in SU
latency	Previous: Time from end of SU to target start. Next: Time from end of target to start of SU.
lcoeff.f0.[p|n]500.[1-5]	Legendre coefficients for F0 contour in previous/next max (500 ms, length of SU) (other speaker)
lcoeff.i0.[p|n]500.[1-5]	Legendre coefficients for Intensity contour for previous/next 500 ms speech
f0diff	|lcoeff.f0.p500.1 − lcoeff.f0.1|
i0diff	|lcoeff.f0.p500.1 − lcoeff.f0.1|

Table 3.8: SWBD-MDE contextual features

towards final rises, greater intensity, longer duration and lower shimmer before *affirmative* backchannels. However, Levow et al. (2010) find that decreased mean pitch and intensity on the last word of the utterance as a cue for feedback in a storytelling dialogue (i.e. a less collaboration dependent task). So, we would like to look at potential contributions of prosodic features of the surrounding speech.

3.4.1 Data and Method

Table 3.8 outlines the contextual features extracted for each of the *really* target utterances from the MDE set. In terms of prosodic features, we model F0 and intensity over the previous 500 ms of speech from the other speaker, and the following 500 ms after the target from whoever speaks the next word. We also identify the last SU that completes before the onset of the *really* (psu) and the first SU that starts after the end of the target (nsu). In some cases the *really* target is produced in overlap, so there are two more relevant SUs: the SU of the last word that starts before the target (pwsu) and the SU of the first word that starts after the target (nwsu). These last two were the same in 118 cases, indicating complete overlap SU wise. There were two cases where all four of these SUs were different, as in the following.

(11) (sw3031)

A: 295.073-296.539 and i just got done with dinner (statement)
A: 296.539-304.279 so um we had beef ribs and a potato casserole and a salad (statement)
B: 296.839-296.949 oh (backchannel)
B: 296.949-297.289 really (backchannel)
B: 297.288-298.225 what did you fix (question)

Our main interest in these SUs is their DA status and timing (i.e. latency) with respect to the target. We are also interested in whether or not the next word is spoken by the same speaker or not (ts: stay or switch). Intuitively, we would expect that stay transitions would indicate a more active use of *really* (although the previous example might suggest otherwise). Similarly, we mark whether the next word is an evaluation by the other speaker (evals). The following gives an example of a negative evaluation by speaker B after the *really* by speaker A, suggesting that the cue word was doing something more than just marking attention.

(12) (sw3901)

B: so i guess it is new in some way although we're learning it's falling apart
A: oh really
B: no[11]
B: that's exaggeration

Our first task is to see whether adding additional contextual information can shed some light on when *really*s get the question label. We do this by extending the multilevel logistic regression from Section 3.3.3. Additional features are added at the individual level (fixed effects) except for the DAs associated with previous and next SUs. The results are given in the following section.

3.4.2 Results

The parameter estimates fitted from the multilevel logistic regression are shown in Figure 3.16. The model intercept term was estimated at -3.65 (se=1.21), which biases predictions towards the backchannel label. Against this, having an evaluative response (evals) has a large effect: it increases the probability of a question label by about 47% (est. 1.86). On the other hand, having the other speaker take the floor in general on the next SU (ts.nsu, possibly after), reduces the probability of a question label by about 11%. Figure 3.17 shows the proportions of backchannels and question

[11]Note: this isn't exactly a passive contribution but it's labelled as a backchannel!

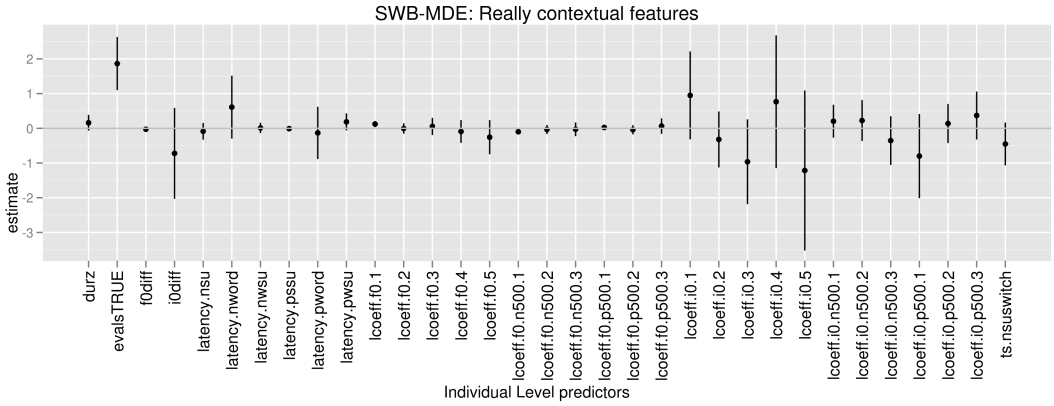


Figure 3.16: SWBD-MDE context: Logistic regression parameter estimates

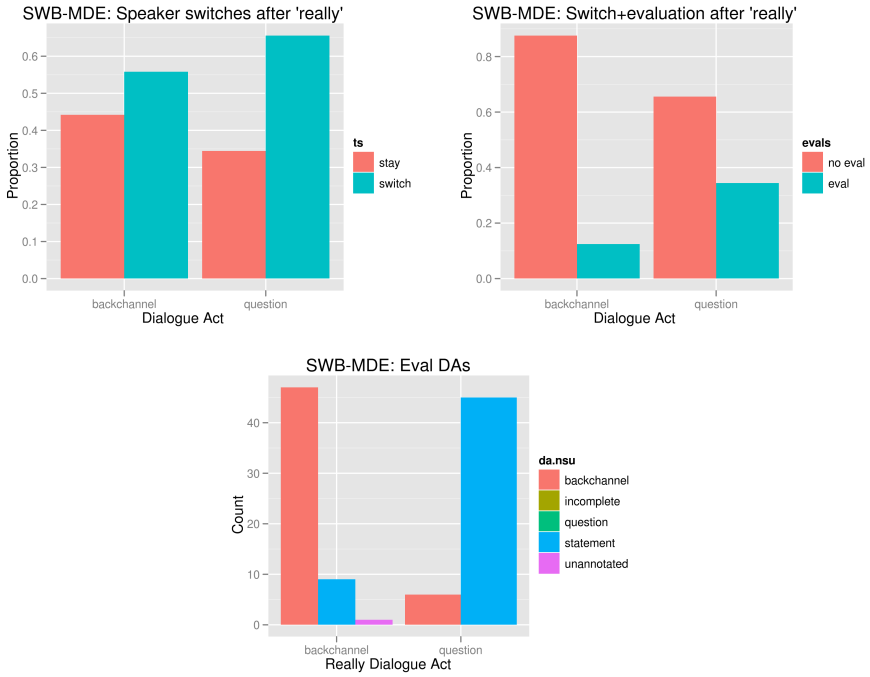


Figure 3.17: *Really* in SWB-MDE: Speaker Switches, evaluations

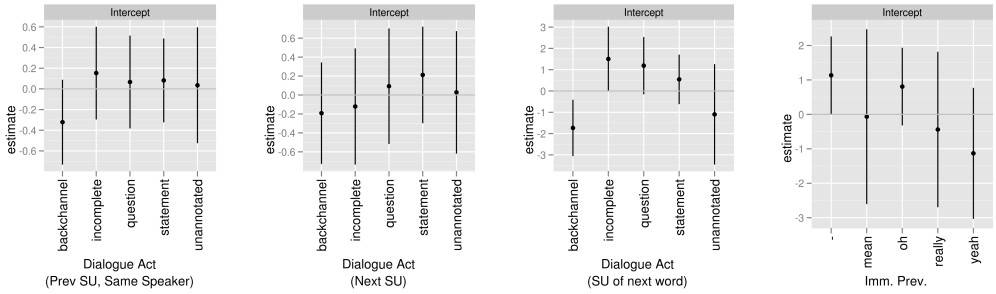


Figure 3.18: SWBD-MDE Context: Parameter estimates for group level predictors.

*really*s that have a speaker switch and evaluatives after them, which illustrates the general trend.

With respect to the prosodic features, we still see significant effects for F0 height (lcoeff.f0.1), from which we get about a 10% increase at the median value (3.34). However, we no longer see a significant effect for intensity concavity (lcoeff.i0.3), although the estimated sign of the effect is still the same. From the contextual features, we see a similarly small but significant negative effect of F0 height in the following 500 seconds (lcoeff.f0.n500.1): every 1 unit higher decreases the probability of a question label by about 2.5%. This low F0 height seems to be a property of the evaluatives, which are on average lower than non-evaluatives in general (mean: 0.10 v 2.01 for eval/non-eval resp.) We don't see any signficant effects from the F0 or intensity features from the previous 500 ms. Similarly, the signs of the latency effects were not definitive, although the general trend was for greater latency to the next word (latency.nword) increased the probability of a question label, and more overlap with the previous turn (negative latency.pwsu) decreased this probability. That is, question *really*s seem to provide more of an interrupt.

Figure 3.18 shows estimates for dialogue acts associated with the surrounding SUs, and the immediate previous cue word. *Really* is more likely to be labelled a backchannel when the SU of the next word is a backchannel and, separately, when the previous turn by the *really* speaker is also a backchannel. The latter estimate suggests that the annotator takes longer distance structure into some account. For example, if dialogue is generally one-sided, an annotator may be more likely to take a string of cue words as backchannels. The estimates for the next SU DAs reflects the fact evaluatives are generally marked as backchannels after backchannel *really* and as statements after question *really* (cf. Figure 3.17). So, it is not clear which comes first. It is certainly not the case that having an evaluative triggers a question label every time. However, the results overall fit with the idea that a question label is applied

when the cue word makes more of an impact in the local discourse. The annotator may use several criteria for judging this. The ones we have seen in these studies for question *really*s were (1) more effortful production of the *really*, (2) the cue word generating an evaluative response, (3) less overlap with the previous SU (although this is just a trend).

3.5 General Discussion

3.5.1 Active/Passive Backchannel Questions

The strongest individual indicator of questionhood appears to be the presence of an evaluative response. At this point we do not have any clear notion of why some evaluatives following *really* should be classified as statements and some as backchannels, so in a way this just puts the problem in another box. One avenue of further investigation might then be to examine the labelling problem for the post-*really* evaluations. However, in doing this we have to deal with the somewhat odd idea that utterances which provoked an evaluation could be considered a passive contribution to the discourse.

We can tease apart some of these issues if we look at bit closer at what cue word *really*s contributes to a dialogue. The crucial point is that *really* seems to be underlyingly interrogative: *really* responding to p is just an elided form of the polar question of the form 'Is it really the case that p'.[12] That is, *really* acts as a check/clarification type move. This underlying interrogativity is reflected in the fact that the majority class for the SWBD-DAMSL set was backchannel question. However, closer inspection suggests that there is more to this class than just a passive signal of attention. This category represents questions about something that is already in the addressee's public commitments. The expectation is that the addressee would answer in the affirmative (though clearly revisions are possible). The basic function of this type of move seems to be to mark contributions to the discourse as new or noteworthy. This differs from affirmatives which seem to mark new additions as congruent with the participants current views. They also differ from true clarification questions which must be resolved for the dialogue to proceed, e.g. correcting misheard words. So, even though *really* is underlyingly an interrogative, it is also conventionalized to the point where not answering it is acceptable, because the *really* addressee has already acted as the source for the proposition in question. We can then take the different types of cue word responses to broadly reflect different aspects of Grice's Maxims. Backchannel questions, like *really*, signal informativity, while affirmatives reflect quality.

[12] Usually actually, 'Did you really?'

From this point of view, it makes sense that the distinction between the backchannel and question categories would be somewhat problematic in the MDE set. The *really*s don't quite fit the canonical idea of backchannel or question. Even though the backchannel question category is available in the SWBD-DAMSL annotation, the results do not indicate that there isn't a uniform notion of what is means to have question form, hence the large number of backchannel labelled tokens. It seems quite likely that annotators would use different strategies for deciding on whether the utterance was an active or passive contribution (similarly, whether something has question form). For example, some might attend more to the overall flow of the dialogue, some might use the presence of an evaluative response to make the decision. At an extreme end, it is possible that some annotators assumed that all *really*s were to be labelled as backchannels. Since it doesn't make much sense for a statement to be a response to a backchannel, this would lead to evaluative responses also being labelled as backchannels. However, we can't distinguish different strategies without knowing who annotated what.

This view fits with the fact that, on the one hand, none of the features investigated above definitively identifies questions or backchannels. However, on the other hand, it does seem that having more of these 'active' features adds to the likelihood that we will see a question label in the annotation. So, while dialogue act classification doesn't seem to be the right dimension to really investigate what the contribution of prosodic form is to interpretation, the data presented here suggest that prosody does make a contribution with respect to whether a response is taken as an active or passive contribution to the dialogue. Questions are active contributions in that they give a task to the addressee. For example, in Farkas and Bruce (2010) the answering task is formulated as picking between different versions of the common ground. Another way to formulate the passiveness pervasive in backchannels descriptions is that they don't put any such requirement on the addressee.

Backchannel questions as a group seem to be somewhere between active and passive in what they contribute. They seem to indicate that previously proffered content in some sense noteworthy by suggesting that the information is worthy of a check. However, this is conventionalized to the point where the addressee doesn't actually have to put much effort into the clarification, and silence will often suffice as a signal of assent. However, it seems that if the *really* speaker puts a lot of effort the production, or immediately follows with a specific question, then the addressee is more likely to act as if there were an actual check - the backchannel question reclaims its 'active' question nature. Nonetheless, it seems plausible that these activeness signals might be mitigated by higher level discourse properties. For example, sometimes it isn't critical that the common ground be of a high quality in how precisely it reflects how things are, e.g. exaggeration is sometimes expected. In such cases, the mean-

ings of these discourse responses tend to become somewhat bleached. In other cases, participants are expected to take the meaning seriously. So, in task-oriented speech we might expect *really*s to be interpreted as real check moves more often.

3.5.2 About those rises...

One working theory then is that high effort features on a cue word (with a conventionalized discourse meaning) will intensify the underlying meaning of that discourse marker. However, this dimension of interpretation seems somewhat orthogonal to the contribution of terminal rises. So what do they do? The key point to remember is that even though we analyze *really* as an interrogative underlyingly, it does seem to have conventionalized backchannel type use in that it doesn't always get a response. So, while it seems that backchannel/question labelling may not precisely represent what *really* does in a dialogue, the results of the corpus analysis are still helpful for evaluating existing accounts of the meaning of terminal rises.

Take the analysis of terminal rises as signalling forward dependency (Pierrehumbert and Hirschberg, 1990). This says that the content in the scope of the rise should be interpreted with respect to subsequent utterances. Questions clearly exhibit this, but this does not really seem to match our idea of what backchannels do. As the check requirement gets relaxed, backchannels questions allow the other participants to move the topic within the limits of relevance. The following gives an example of this sort of use.

(13) B: Oh I've got some Chinese Hollies that are just outrageous
B: They they are very sharp
A: Oh really?
B: Do you do your own uh lawn maintenance?
A: Yeah

The idea that falls mark speaker responsibility, along the lines of Steedman (2000), is hard to fit with the fact that we get question *really*s that are falling. In the same vein, accounts which suggest that rises add a level of contingency (Gunlogson, 2008) or a test on the common ground (Nilsenova, 2006) with respect to the proposition under the scope of the rise are somewhat problematic, given that the majority of *really*s don't get an evaluative response with or without a rise. The way around this is to say that *really*s with rises are contingent/testing but only weakly so, and not any more than the falling type. However, this doesn't solve the problem of what we do with affirmative backchannels which can also rise (Gravano, 2009).

What is common to backchannels and questions is that they are discourse medial. That is, there is some expectation that the other participant will take the floor afterwards. However, this doesn't necessarily mean that the cue word content somehow shouldn't be attributed to the cue word speaker. So, the fact that some questions fall and some backchannels rise fits with the idea the idea that rises signal discourse openness (Cruttenden, 1997) or non-finality (Bartels, 1999), while falls are underspecified. The distribution of rises in the data then leads us to an analysis where rises appear probabilistically at 'open' points in the discourse. The probability of getting a rise then broadly depends on how important it is for the signal of non-finality to come through. This sort of account is one of exclusion as suggest in (Kowtko, 1997): rises generate extra meaning at points where closure is expected. Putting a rise on *really* is not so important because of its inherent interrogative nature. However, putting a rise onto the conversational closer in (14) adds a questioning feeling. Similarly, adding a rise to a strong affirmative like *right* seems to work naturally when that the current question/task under discussion has not been resolved.

(14) A: Goodbye!
B: Goodbye?

(15) A: How do I get the thread on the bobbin?
B: You put it on the little bar on top of the machine
A: Right...
B: And then...

So, it seems that any uncertainty on affirmatives added by rises is directed at whether the last offering resolves the QUD rather than whether the proferred content is true. The fact that affirmative backchannels are characterized by rises in the task-oriented dialogues suggests that the distinction between backchannels and agreements is really about whether the speaker seems to expect a continuation from the other speaker or not, and whether speaker deems it necessary to signal this, rather between levels of grounding.

3.5.3 Dimensions of Prosodic meaning

From a synthesis point of view, if we wanted to make the questioning/check nature of *really* more salient, we would put more effortful features on it: we would set a high mean F0, intensity and add an F0 rise. But in order to understand if there are conventionalized meanings associated with pitch contour shape, we need to look elsewhere. The corpus studies presented above clearly show a lot of variation in the *really*s in terms of F0 shape. It appears that this variation is only indirectly

represented in dialogue acts, so the question is what other dimensions of meaning can capture this variation, if any? The discussion above leads us to two dimensions of meaning upon which to examine the contribution of prosody: how active the underlying meaning of the cue word is, and what state does the utterance leave the dialogue, open or closed. We expect that we will find a more direct effect of terminal rises on the latter.

If variation in these dimensions are not really illuminated by current dialogue act labels, then how should we go about investigating their relationship with prosody? The obvious place to look is in attitudinal correlates of prosodic variation. For example, assuming that the discourse meaning of cue word *really* is to express surprise or noteworthiness, then we would expect having more effortful features would increase the perceived level of surprise. Note, however, that under this view this expression of propositional attitude derives from semantics of the cue word rather than the prosodic form, so that the effortful prosody acts more like other lexical intensifiers, e.g. *very*. This separation of underlying meaning from the contribution of prosody differs from accounts of rises which have associated rises directly with attitudes around uncertainty and submissiveness. Again, we come up against the question of what it would mean for a backchannel, particular an affirmative, to express uncertainty. The main distinction to make is between whether rises express some sort of structural rather than propositional uncertainty as suggested above (cf. Example (15)), i.e. that rises signal that the QUD is unresolved. We take up the perception of attitude and cue words in a series of perception experiments in the next chapter.

3.6 Conclusion

In this chapter we examined the meanings expressed by cue words in terms of dialogue acts. The point of these empirical studies was to understand how prosody affected the interpretation of these responses. In this way, this work differs from the previous studies in that we look at how dialogue acts can be used to help us understand the role of prosody, rather than the other way around. The first pass hypothesis was that specific prosodic cues associate with different dialogue acts. Following the literature we would expect to associate uncertainty and questioning with terminal pitch rises. This corpus study of *really* was an attempt to find out if rises, or some other prosodic feature, could systematically differentiate question *really* and backchannel *really* in the similar manner to the way final rises have been argued to change the interpretation of declarative sentences. It seems clear at this point that the prosodic features considered are not enough to make this distinction. We can get better than baseline accuracy in the classification task, when we train models on the downsampled data. Nevertheless,

a lot of backchannels go up and questions go down. This suggests that a good deal what makes something questioning is in the context and if prosody does help to distinguish the type of act, it only does so indirectly. Instead prosody seems to convey more general meaning which composes with the semantics and pragmatics of the cue word which can lead to a questioning interpretation. Combining the findings of the current study and previous investigations of affirmative cue words, the account of rise meaning that best fits the data is the very general non-finality one.

So, what does the variation in prosody we have observed actually mean for the interpretation problem? To understand what is going on we want to look at attitudinal variables. For the case of *really*, the intuition was that we feel like we make big *really*s when someone has added something surprising to the discourse. That is, something worthy of a check. It seems then that prosody can act on multiple dimensions. On the one hand, prosodic features can act to intensify the underlying meaning (Bolinger, 1972), or increase general saliency (which then is perceived as importance - 'I care that this goes in'). On the other hand other conventionalized cues like rises in English can be modelled as signalling something about the current discourse configuration, while not necessarily signaling a specific act. We take up the question of how these two dimensions of prosodic variation affect interpretation in the next chapter.

Chapter 4

The Interpretation of Cue Words and Rises

4.1 Introduction

The previous chapter investigated potential links between prosodic form and dialogue act tags in the Switchboard corpus. The results of that study suggested that prosody only had an indirect contribution to this sort of discourse categorization for cue words. Moreover, it seemed that there were at least two dimension of variation at play with respect to discourse meaning. We hypothesized, on the one hand, that more effortful features – higher mean pitch and intensity – should intensify the underlying meaning of the cue word. On the other hand, terminal pitch rises signal that the dialogue is in some sense incomplete. To test these hypotheses we need response variables other than the dialogue acts examined previously. In fact, it seems that the best approach for teasing apart the contributions of cue word semantics and prosodic features is to use attitudinal measures.

As such, this chapter presents two experiments investigating how perception of speaker attitude varies with prosody. I previously argued that the main use of cue word *really* is to indicate surprise. So, the first experiment investigates how the perception of surprise varies with the prosodic features of this cue word, as well as one with very different semantics: the affirmative *right*. To tie things back to the previous studies about dialogue acts, we also look at how questioning these cue words are perceived to be given a variety of prosodic forms. The effect of the underlying semantics is further investigated in the second experiment, where we look at the relationship between terminal rises and the perception of uncertainty. In particular, we test the hypothesis that final rise signal that the QUD is unresolved. That is, the

type of uncertainty signalled by the rise is directed at the discourse structure, rather than expressing a propositional attitude.

In the end, the results of these experiments support our initial hypotheses. It seems that having more effortful features correlate with a greater perception of surprise and questioning for *really*, but not for *right*. Adding a rise to a cue word increases the perception that more needs to be said on the current topic. However, the underlying semantics of the cue word dominates the perception of the speaker's attitude toward the proposition under discussion at that point. These results give us a clearer understanding of what these cue words and prosody do in dialogue and how these moves fit in with semantic/pragmatic theory more generally. Moreover, this sheds light on why the distribution of cue words differs in corpora representing different speech tasks. Examining how the cue words fit into into extant semantic/pragmatic theories also helps us get a grip on the notion that standards for common ground update can change with the context, and more generally the managing of the concept of gradability at the propositional level. This also allows us to make distinctions between the intensification effects produced at the semantic level, and those produced prosodically.

4.2 Perception Experiment: Cue Words and Speaker Attitude

The corpus studies in Chapter 3 found the prosodic features considered (which included various pitch, intensity and duration features) could not be used to separate the data into MDE backchannel/question annotation set, suggesting we needed to consider other response variables for prosodic variation on *really*. This section presents an experiment investigating the perception of *really* and *right* in terms of how surprising and how much like a question they sound. From this study we find pitch range to be the feature best correlated with perceived surprise. We also find the question and surprise ratings to be correlated.

4.2.1 Data and Method

The stimuli were selected from the MDE data set analyzed in Chapter 3. The stimuli set consisted of 192 tokens taken from three subsets: backchannel *really*s (MDE labelled), question *really*s, and *rights*. Each subset consisted of 64 tokens selected to represent the different features. The set of *really*s examined in the previous corpus study was split based on quartiles based on pitch range, pitch level and duration

(the *rights* were processed in the same way). Pitch level is an indicator of the height of the utterance relative to the range of values exhibited by the speaker in the rest of the conversation. Pitch measurements at the 1st and 99th quantiles averaged to approximate pitch range extremes for each speaker (nmin, nmax). *Pitch level* was approximated as the (pmax - nmin)/(nmax - nmin) where *pmax* was the maximum F0 value of the *really*. The data was first split into four groups according to *pitch range*. These groups were then split into a further four groups according to *pitch level*, and the similarly *duration*. One stimulus was then randomly selected from each of the the 64 groupings. Eight University of Pennsylvania affiliates (5 female, 3 male) participated in this study. All subjects were native speakers of English and they had an average age of 23 years. The subjects were paid to participate in this experiment.

The randomized stimuli were presented via a computer interface. The subjects listened to each stimuli through headphones and were allowed to replay the the current stimuli as many times as they liked. Subjects were asked two questions with respect to each stimuli: 'How surprised does the speaker sound?' and 'How much like a real question does this sound like?'. They were then directed to answer these questions on two 7 point sliding scales (1=not at all, 7=extremely). The subjects were given a chance to ask questions and confirmed that they understood the task.

4.2.2 Results

The average rating for surprise versus question for the stimuli are shown in Figure 4.1. This shows the correlation between these two ratings (Kendall's $\tau = 0.63, p < 0.001$, distributions are non-normal).[1] The figure also shows a lack of association between either rating and the MDE backchannel/question annotation. The ratings for backchannel/question categories are not significantly different (Mann-Whitney U test: question $p = 0.30$, surprise $p = 0.18$).

Generally, subjects appeared to find the lexical constraint quite strong. None of the *right* stimuli had an average question rating above 4 (the midpoint on the scale). Thus, lexical constraints interacted with how the prosodic cues were interpreted. Along the same lines, subjects did not behave completely uniformly in rating the stimuli. Subject variation and prosodic features are discussed in the following subsections.

Really	τ_q	p-value	τ_s	p-value
pitch range	0.533	0.000	0.581	0.000
pr1	0.339	0.000	0.426	0.000
pr2	0.451	0.000	0.497	0.000
pitch level	0.414	0.000	0.502	0.000
slope	0.172	0.005	0.161	0.008
slope1	0.428	0.000	0.504	0.000
slope2	0.005	0.931	−0.035	0.567
duration	0.285	0.000	0.254	0.000
d1	0.216	0.000	0.230	0.000
d2	0.278	0.000	0.225	0.000
intensity	0.130	0.033	0.272	0.000
Right				
pitch range	0.240	0.007	0.285	0.001
pitch level	0.111	0.210	0.278	0.002
slope	0.234	0.008	0.093	0.299
duration	0.162	0.066	0.154	0.084
intensity	0.198	0.025	0.374	0.000

Table 4.1: Correlation coefficient (Kendall's τ) and p-values of the question/surprise ratings and prosodic features for *really* (top) and *right*

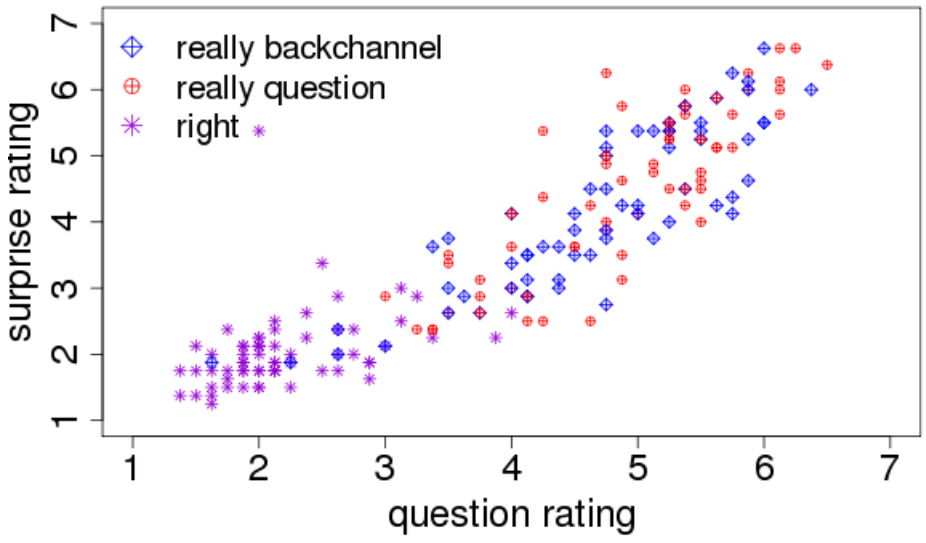


Figure 4.1: Average surprise versus question ratings.

4.2.3 Prosodic Features

Table 4.1 show the correlation (Kendall's τ) between the ratings and various prosodic features for the *really*s (overall and by syllable) and the *rights*. We find the question/surprise ratings to be most highly correlated with pitch range and pitch level. This fits in with an *effort code* style interpretation: more effortful prosody is associated with more perception of surprise (Gussenhoven, 2002).

Somewhat unexpectedly, the second syllable slope of *really* was not significantly correlated with either questioning or surprise while the first syllable slope was correlated with the ratings. That is, the final fall/rise does not seem to contribute to whether the *really* is interpreted as a question. This is not to say that final rise/fall never contributes to question interpretation. However, in the case of *really*, it does not add much to this dimension in the face of expanded pitch range. It is also interesting to note that mean intensity of the first syllable does not with the syllable pitch range ($\tau = 0.04, p = 0.62$). So, surprise value of *really* seems primarily signalled by the size of the pitch excursion in the stressed syllable. We also note that the size of

[1]The correlation coefficient τ is interpretable as the probability that the observed data are concordant (same sign, going in the same direction) minus the probability that they are discordant (different sign).

Subject	mean_q	sd_q	mean_s	sd_s
1	3.49	1.93	2.97	1.83
2	4.39	1.85	3.98	1.73
3	4.14	1.11	4.06	1.56
4	2.39	1.60	2.11	1.32
5	4.64	1.86	3.88	1.85
6	4.31	1.84	3.91	1.76
7	4.11	2.54	3.52	2.40
8	3.94	1.64	4.19	1.57

Table 4.2: *Means and standard deviations for question and surprise ratings, by subject*

this pitch excursion on the first syllable is not highly correlated with that syllable's duration ($\tau = 0.19, p = 0.001$).

In general, it seems that the mapping from prosodic feature values to question levels do not match between *really* and *right*. For example, the mean question rating for *really*s with pitch range between 5 and 10 semitones is 4.93, while the corresponding value for *rights* is only 2.41. However, the data shows that these 'barriers' can still in fact be passed for *right*. One *right* stimulus received a relatively high average surprise rating of 5.375. In fact, this stimulus had the highest ranking pitch range, pitch level and duration of the *right* set (5.68, 3.08 and 2.49 standard deviations from respective means). Note, however, that the same stimulus still received a low question rating (2.0). So, for *right*, expanded pitch range, level and duration may contribute to the perception of surprise independent of questionhood.

This leads to the question of whether rising intonation can be interpreted as questioning with an agreement particle like *right*. Of the 64 *right* stimuli, 33 had F0 contours with overall rising slope. In fact, all the stimuli with average question rating greater than 3 have positive slope (6/33 items). However, most rising *rights* did not sound questioning to the subjects in terms of which side of scale ratings fell. Inspection of one pair of stimuli with similar features, shows that closer fitting of F0 than simple linear regression may be necessary to sort this out. For example, the lower rated item had a perceptible final fall even though the general trend was positive. We will return to question of rises and *right* in the Section 4.3.

4.2.4 Subject Variation

We also need to consider how well these results apply to the range of subjects. Using Krippendorff's α for ordinal data (Krippendorff, 2004, Artstein and Poesio, 2008),

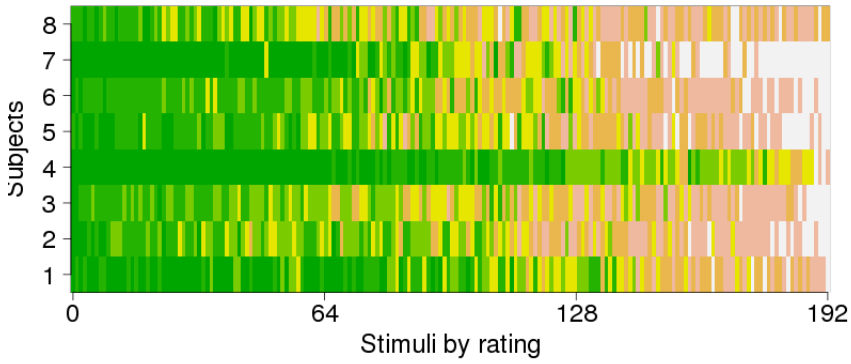


Figure 4.2: Stimuli ratings ordered by mean average rating (increasing rightwards) by subject. Subject 4 was significantly different from the rest (Pairwise U tests: $p < 0.001$).

	1	2	3	4	5	6	7
2	***						
3	0.18	0.16					
4	***	***	***				
5	***	1.00	**	***			
6	***	1.00	0.81	***	1.00		
7	0.12	1.00	1.00	***	1.00	1.00	
8	0.74	0.09	1.00	***	**	0.35	1.00

Table 4.3: *Pairwise Mann-Whitney U tests for question rating by subject with Bonferroni correction (** = $p < 0.01$, *** = $p < 0.001$).*

we see that agreement between raters was above chance but still not extremely high ($\alpha_s = 0.58$, $\alpha_q = 0.50$). Closer examination of the data suggests that subjects had different rating biases. Table 4.2 shows means and standard deviations of the ratings by subject. The responses of subjects 1 and 4, in particular, appear significantly lower than the other participants. This is confirmed via pairwise Mann-Whitney U tests (Table 4.3). However, these subjects did not simply avoid the higher end of the scale. Subject 4, for example, did give stimuli with extreme pitch range values similarly extreme surprise ratings.

Figure 4.3 shows the distribution of pitch range versus first syllable slope for the *really* data used in the experiment. It also highlights the stimuli which were high (> 5) and low (< 3) rated as questions by subject 4. In this case, the ratings seem

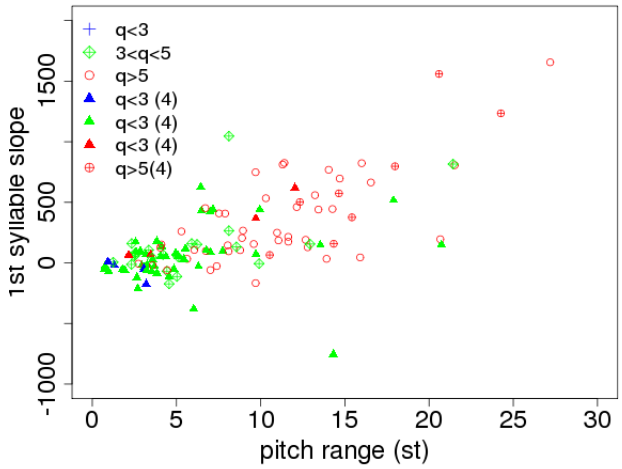


Figure 4.3: Pitch range versus first syllable slope. The colors indicate different average ratings. Points highlight the ratings of our most conservative subject (4).

based on the same prosodic inputs as for the other subjects but the rating cutoff points appear to be higher.

4.2.5 Discussion and Summary

This section investigated how prosody affects the perception of two cue words with very distinct meanings. *Really* and *right* appear to induce different mappings from prosodic features to surprise/question ratings. This supports the hypothesis that effortful features emphasize the underlying meanings of these words. Since *right* doesn't inherently express unexpectedness or surprise, it makes sense that it scores lower than *really* on this dimension for similar prosodic features. It seems that the addition of effortful features, like expanded pitch range, appear to bring the interrogative nature of *really* to the forefront.

Given the correlation between the surprise and questioning ratings, it is not actually clear which leads to which. *Really* might be perceived as expressing surprise because it reaches some threshold of active questioning, or vice versa, where these thresholds are still somewhat variable from speaker to speaker. Nevertheless, we see that the underlying meaning of *really* appeared to be more salient to how questioning the utterance was perceived than the final fall/rise characteristics. Hence, second syllable slope did not turn out to be a good predictor of the question rating.

The question ratings in this experiment did not match the SWBD-MDE question annotations of *really*. This does not mean that these dialogue act type categories are not meaningful. Instead, it highlights the indirectness in the mapping between prosody and dialogue acts. Recall, also that SWBD-MDE annotators had the full textual and audio context to consider while performing the labelling, while in the experiment the subjects really only had the prosody to work with. So, what we see is that the level of surprise/questioning is a good measure of prosodic variability for *really*, but not so much for *right*. Again, we see a lack of parallelism between questions and rise/fall characteristic. However, rises still do seem to do something if indirectly. Even though *right* sat low on the question scale, we did see a significant positive correlation between question ratings and F0 slope, though this did not translate to the surprise rating in this case. So, we would still like to better understand what adding a rise to an affirmative does. In the next experiment, we look more closely at the contribution of terminal pitch rises to the interpretation of a range of cue words.

4.3 Perception Experiment: Terminal Rises and Uncertainty

The previous experiment supported the idea that more effortful prosodic features highlighted the underlying attitude of cue words. For *really* this correlated with the perception of speaker surprise. However, this did not clear up the issue of what rises contribute in these situations. Final rises have been associated with speaker uncertainty in both theoretical (Gussenhoven, 2002, Nilsenova, 2006, Reese, 2007) and empirical studies (Pon-Barry, 2008, Gravano et al., 2008, Litman et al., 2009). So, this seems a good place to look for an attitudinal correlate of rises. However, as discussed in the the previous chapter, the idea that rises contribute a signal of uncertainty doesn't quite fit observed the prosody of affirmative cue words, which has been found to be predominantly rising in backchannel and continuer type moves (Benus et al., 2007, Ward and Escalante-Ruiz, 2009).

Being able to detect attitudes like uncertainty is clearly important for developing dialogue models from both a theoretical and practical standpoint. So, we would like to find out if rises really are a consistent indicator of this sort of attitude. One reason that we get these seemingly different uses of rises is that the uncertainty expressed can be directed at different parts of the discourse. In some cases the uncertainty may be about the semantic content expressed in the utterance itself (propositional uncertainty). A prime example of this is the rising declarative question. However, it seems that uncertainty can also be directed at, for example, the QUD (discourse structural uncertainty). In particular, I suggested that rising affirmatives signal that,

while the speaker agrees with the previous utterance, they might be uncertain as to whether it resolves the current QUD. These sorts of distinctions are not usually spelled out in investigations using uncertainty as a response variable, but they clearly make a difference to how a discourse should proceed.

As such, the current study investigates how rising intonation interacts with six cue words with a a range of discourse meanings. At one end of the spectrum, affirmatives like *right* and *yeah* primarily express agreement with the utterances they respond to. Other affirmatives like *okay* and *sure* express acceptance of a request, which may simply be to accept last utterance in the dialogue. As such, they do not seem to express as strong agreement as *right*. At other end, responses like *really* and *well* generally express an inability or unwillingness to admit the utterance at issue into the common ground (Schiffrin, 1994). We use pitch resynthesis to look at how variation in intonation relates to cue word speaker's attitude to the propositional content they are responding to (the *at-issue content*). We look at this problem via attitude scales which touch on different aspects of certainty/uncertainty in dialogue: certainty about whether the proffered content is true/acceptable, whether the proffered content is unexpected, and whether the matter requires more discussion. The last QUD related notion basically reflects the idea that rises signal that the discourse has not come to a good stopping point.

4.3.1 Data

The stimuli for for this experiment were once again drawn from the Switchboard corpus. Each stimulus consisted of a (textual) context and (audio) cue word response. Since the goal was to see if different types of rises and falls would affect cue word interpretation, the rises and falls were varied in terms of the height of a peak/valley and overall pitch range using resynthesis (PSOLA in Praat). From the previous experiment we expected that stimuli with larger pitch range would signal greater surprise, while higher peaks/lower valley would produce more emphatic interpretations. Context types were chosen to represent different levels of certainty. The goal here was to test whether lexical markers of uncertainty in the context would be mirrored in interpretation of the response.

Cue words. The resynthesized responses were derived from 6 base cue words: *really*, *well*, *okay*, *sure*, *yeah*, and *right*. Two tokens of each base word were randomly selected for resynthesis. Base tokens were drawn from occurrences of the cue words in one word turns according to the transcripts, and were checked for modal voice quality. Resynthesized contours were set with respect to the start, end, and the midpoint of the stressed vowel (nucleus for diphthongs). F_0 values for the stylized contours were based on quantiles derived from F_0 values from other turns of that speaker in the same

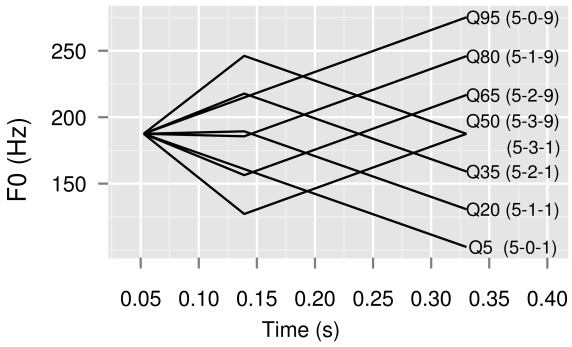


Figure 4.4: Stylized pitch contours for *right*, with quantiles and contour mnemonics.

conversation. Each base token was resynthesized in 8 ways as shown in Figure 4.4, so that the start point was always the median value and the gradient between the mid- and endpoints remained the same. So, the stimuli in each group varied in pitch range but maintained the same slope at the end of the word. The stimuli were resynthesized using PSOLA via Praat. The resynthesized versions were checked for naturalness and that each contour was audibly different.

Contexts. The contexts were drawn from turns that occurred immediately prior to one of the cue words. Four different types of context were selected for each cue word. As mentioned previously, these types were chosen to represent different levels of certainty, although they clearly do not exhaust the possible categories. Four types of context were used: (i) factual, e.g *X is Y*, (ii) evaluative, e.g. *X is good*, (iii) attributed, e.g. *I heard that X*, (iv) inferred e.g. *probably X*. Four turns were selected for each context type. So, the stimuli consisted of 6x2x8 = 96 cue words and 6x4x4 = 96 contexts in total.

4.3.2 Method

14 native speakers of American English, undergraduate students, participated in this experiment. Subjects were paid for their participation. The experiment was presented via a web interface formulated using WebExp.[2] Subjects were told that they were going to be presented with snippets from real telephone conversations. They were presented with a statement and cue word response which they could listen to as many times as they chose by pressing a button. Contexts and responses were randomly paired. Subjects were asked to provide ratings on a 1-7 scale as answers to the following questions:

[2] `http://www.hcrc.ed.ac.uk/web_exp/`

1. How expected does what A said seem to B?
   (1=completely unexpected, 7=completely expected)
2. How credible does what A said seem to B?
   (1=not at all credible, 7=completely credible)
3. Given B's reaction, how much would you expect A to explain or provide more evidence for what they say/why they said it?
   (1=wouldn't expect a follow up, 7=definitely expect a follow up).

Rather than ask directly about uncertainty, the idea was to relate uncertainty to different aspects of dialogue structure. Question 1 (EXPECTEDNESS) reflects certainty with respect to B's prior beliefs. Question 2 (CREDIBILITY) reflects how willing B is to believe A, i.e. add the content of A's utterance to their public beliefs. Question 3 (EVIDENCE) reflects the status of the QUD, i.e. whether A's utterance has been resolved/accepted or whether it is still contentious.

Subjects first completed 4 practice slides to familiarize themselves with the task. All participants reported that they understood the task before moving on to the main experiment, which consisted of 64 more slides in the same format.[3]

4.3.3 Results

General Trends

The mean scores for each cue word, grouped by question, are shown in Figure 4.5. The EVIDENCE scale has been reversed so that low scores indicate a lack of resolution of the question under discussion, hence uncertainty. Scores are generally higher for affirmative cue words with falling intonation over all of the questions. It also appears that scores increase with the affirmative strength of the cue word. With falling intonation, agreement markers *yeah* and *right* seem to convey more certainty than *okay* and *sure*. As expected, *really* and *well*, which mark discord in the dialogue, have lower scores.

On inspection, rising intonation appears to have less of an affect on CREDIBILITY than the EXPECTEDNESS or EVIDENCE scales. In the later two cases, rising intonation pushes scores towards the uncertain end, most strikingly for *yeah*, but also for and *really*, *okay* and *right*. However, this does not seem to be the case for *well*, which seems to have the opposite trend.

[3]Note: due to a calculation error not all contexts and cue words were presented to each subject. However, the unbalanced nature of the data set is not a problem for the multilevel model used to analyze the data in the following subsection.

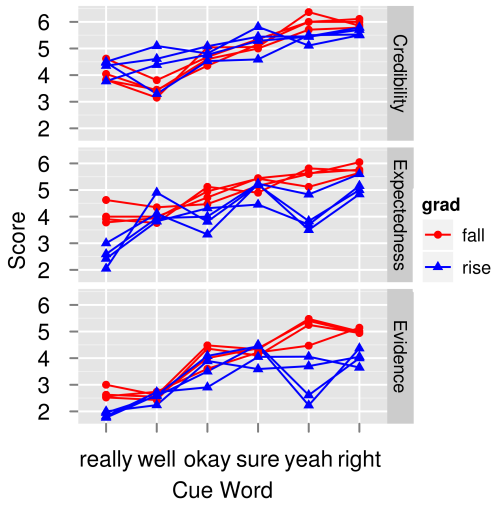


Figure 4.5: Mean scores for each cue word by question (question 3 reversed).

Modelling the Data

A multilevel model was fitted to help sort out the effects of cue word and contour, as well as context and subject variability. Following the approach outlined in Gelman and Hill (2007) (also Chapter 3), observed scores, y, were modelled as follows.

$$(1) \qquad y_i \sim N(\mu + \alpha^{cw}_{j[i]} + \alpha^{ct}_{k[i]} + \alpha^{cx}_{l[i]} + \alpha^{s}_{m[i]} + \alpha^{cw.ct}_{j[i],k[i]}, \sigma^2_y)$$

With group level predictors modelled as follows:

$$\begin{aligned}
\alpha^{cw}_j &\sim N(0, \sigma^2_{cw}) \text{ for } j = 1, \ldots, 6 \\
\alpha^{ct}_k &\sim N(0, \sigma^2_{ct}) \text{ for } k = 1, \ldots, 8 \\
\alpha^{cx}_l &\sim N(0, \sigma^2_{cx}) \text{ for } l = 1, \ldots, 4 \\
\alpha^{s}_m &\sim N(0, \sigma^2_{s}) \text{ for } m = 1, \ldots, 14 \\
\alpha^{cw.ct}_{j,k} &\sim N(0, \sigma^2_{cw.ct}) \text{ for } j = 1, \ldots, 6, k = 1, \ldots, 8
\end{aligned}$$

Here α^{cw}_k is a parameter representing the effect of cue word k holding the other variables constant. Contour (ct), context (cx) and subject (s) and the interaction between cue word and contour (cw.ct) were similarly modelled as separate groups. The coefficients within each group were modelled as arising from different normal distributions, however their means are pulled out into a grand mean μ. The model parameters, along with finite population standard deviations for each group, were

estimated using the Markov Chain Monte Carlo technique as implemented in JAGS[4] via the R package `rjags`. The model estimation passed Gelman-Rubin and Geweke convergence diagnostics.

Parameter Estimates

Figure 4.6 shows estimated medians and 95% intervals for the different parameters for each of the scales. The finite population standard deviations give us a measure of how variation in the actual data is associated with each factor, which basically gives a bayesian version of the classical analysis of variance. We can immediately note that a large source of variation was the subjects themselves. Subjects appeared to have different strategies for the different scales. Abstracting away from this, we can consider estimated effects of cue word, contour and context.

The parameter estimates for context type are very small, and estimates fall well inside the 95% intervals of the other context types and thus do not appear to be significant. So, the interpretation of these responses did not seem to depend on the semantic context types provided in this experiment. This is also reflected in the small finite population standard deviation estimates. The greatest standard deviation is associated with cue word identity. That is, cue word semantics appear to have a large effect on the perception of response credibility, unexpectedness and the need for more evidence. Again, the effect goes in the same direction as the strength of affirmation for all scales. This is not the case for the contour results, where we see a clear distinction between the CREDIBILITY rating and the other scales. For the EXPECTEDNESS and EVIDENCE scales, rising intonation pushes scores towards the low end of the scale. The posteriors associated with falls and rises appear quite distinct with medians for rises generally lying outside the 2.5th quantile of the falling contours. This effect is not present for CREDIBILITY. This again is reflected in the finite population standard deviations for the contour group: the posterior includes zero for the credibility model, but is above zero for the other two ratings.

Figure 4.7 shows the results for the cue word/contour interaction term. We can see that the effect of rising intonation varies across cue words. As in Figure 4.5, the greatest effect appears to be with respect to *yeah*, with rising contours pushing scores downwards for EVIDENCE and EXPECTEDNESS, while falling contours pull the scores up. A similar trend is observed with *really*, although interestingly variation appears to be mostly on the EXPECTEDNESS scale. On the other hand, rising contours appear to raise *well* scores.

Although the general trends for rises and falls seem fairly robust with respect to

[4] `http://www-fis.iarc.fr/~martyn/software/jags/`

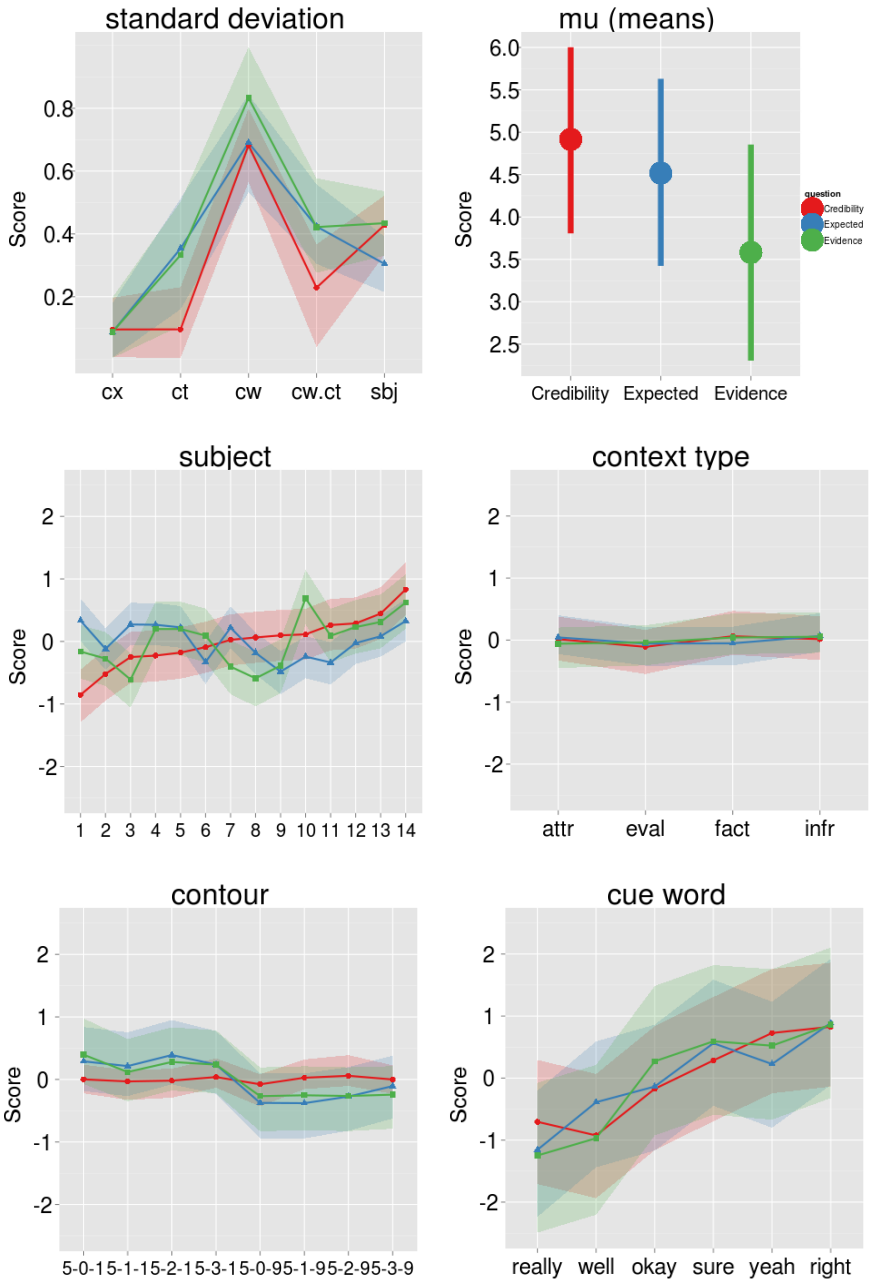


Figure 4.6: Parameter estimation medians. The shaded range represents 2.5th-97.5th quantiles. Red: CREDIBILITY, blue: EXPECTEDNESS, green: EVIDENCE.

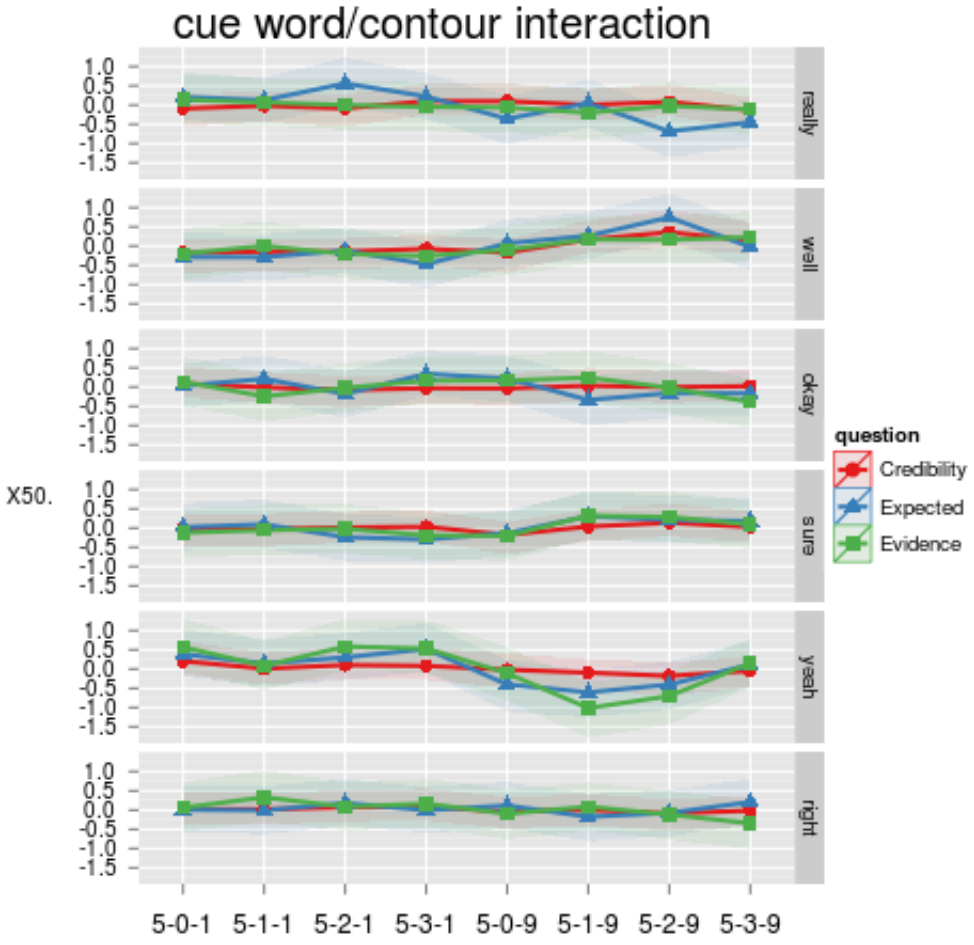


Figure 4.7: Cue word/contour interaction

unexpectedness, we do not see much of a distinction between the different types of falling and rising contours (c.f. Figure 4.6). Greater pitch ranges were not really associated with the perception of more unexpectedness. This is somewhat at odds with the previous results linking pitch range to the perception of surprise presented in the previous experiment (similarly Gussenhoven, 2002). It may be that the connection between pitch range and surprise may be more to do with the velocity of the intonational gesture rather than peak height or overall pitch range. Note, since the resynthesis was based on quantiles, we cannot really draw strong conclusions based on the individual contours across cue words. However, given that the stimuli were generated over 90% of the speaker's pitch range, the general within-word trends seem clear and a more detailed exploration of the interaction of range and rises is left for future work.

4.3.4 Discussion and Summary

The Interpretation of Rises

These results support the hypothesis that that rising intonation signals some sort of incompleteness in the dialogue, rather than a lack of credibility or propositional uncertainty. This credibility/certainty is clearly reflected in the choice of cue word. The fact that intonation did not have much of an effect on the credibility scale suggests that, when rises do map to an uncertain attitude, this reflects difficulty integrating the new information proffered by the other speaker, rather than expressing some sort of uncertainty or disbelief about the content (cf. Shimojima et al., 2002). By signaling that the QUD is unresolved the speaker implicitly signals that resolution depends on the hearer.

For cue words, the inability to resolve the QUD may happen when the utterance under scrutiny does not fit with the responder's beliefs. The content may be epistemically unexpected (i.e. it doesn't fit their world view). However, another possibility is that the content is unexpected from the point of view of relevance. This experiment did not differentiate these two cases. However, the latter case seems to apply to strong agreement words like *right* pronounced with a rise. That is, the responder may agree with the content, while not being quite sure how it relates to the current QUD. So, again, while rises are response evoking in a weak sense, they do not necessarily make an utterance an interrogative.

Rises and cue word semantics

Although *right* has higher scores than *really* on all the scales, we still see similar distinctions between rising and falling contours in terms of EVIDENCE and EXPECTEDNESS scales. However, the interaction with rises sheds light on how cue words with more similar semantics vary in meaning. With respect to the affirmatives, we see that *yeah* appears to be able to express more unexpectedness then *right*. This seems to be attributable to the fact that *right* is a stronger affirmative: it conveys that the respondent believes the content at-issue to be true (hence the high credibility scores) However, while *yeah* conventionally expresses agreement, it does not reveal so much about the responders private beliefs. Thus, *yeah* with rising intonation can be interpreted as conditional acceptance while simultaneously asking for more evidence. This sort of request for more evidence would be pragmatically odd when the speaker is already known to believe the content, as would be the case for *right*. So, in a sense, prosody is able to influence the interpretation of *yeah* more than *right* because its semantics is not as specific.

A similar contrast is evident between *really* and *well*. The latter signals that acceptance of the content under scrutiny is not possible for that speaker given the current state of the dialogue. We saw previously that when *well* is a one-word turn it is usually marked as an incomplete or abandoned turn in terms of dialogue act (cf. Section 3.2.2). That is, *well* is conventionally turn holding. It is plausible that this aspect of *well* overrides the contribution of the rise, so it is a stronger cue word than *really* in this respect. On the other hand, like *yeah*, *really*'s semantics appear to allow for more shades of meaning. Overall, we see that the interpretation of rising intonation with cue words depends on the (discourse) semantics of the cue word. In particular, how much this reveals about the speaker's beliefs.

4.4 Cue words, Gradability and Discourse Structures

The previous two perception experiments indicated that the semantics of the cue word affect how the prosodic features like final rises are interpreted. Increasing the prosodic effort on a cue word seems to highlight the underlying meaning of the cue word, rather than adding an independent attitude like surprise. Similarly, we saw more variation in interpretation in some cue words than others when we manipulated the pitch rise/fall characteristic. In rest of this chapter, section we look further into the meaning of cue words to see why we get the results we do. To get a better idea of how these cue words relate to the discourse structures, we first look at some distributional differences

Corpus	Yeah	Right	Sure	Okay	Really	Well
Let's Go	113	54	0	93	1	30
Columbia Games Corpus	903	189	-	2247	-	-
HCRC Maptask	1642	≈1500	3	2360	0	960
ICSI Meeting	11482	4420	286	4766	218	2499
Switchboard (NXT)	11922	2797	308	1540	535	5364

Table 4.4: Cue word frequencies across corpora. Note, only affirmative counts have been released for the Columbia Games Corpus (Gravano, 2009).

across dialogue styles. We will then take some time examining the meaning of *really*, showing how the cue word version relates to it's sentential use and how this can be analyzed, drawing on existing accounts of gradability and modals. This sets us on the way for similar accounts of affirmatives like *right*.

These investigations bring us to a key part of why discourse needs so much managing in the first place: standards vary from situation to situation. The crucial standard we are dealing with in modelling dialogue is the one that determines whether we admit an informational unit into the common ground. That is, the evaluations of propositions are not really absolute. Instead there seems to be some inherent gradability in this area. Studying how cue words fit into this sheds light on how this gradability relates to the discourse structures we have been dealing with and how this can be modelled. It also helps us understand why certain cue words are favoured in specific situations and, ultimately, how we expect them to combine with other discourse oriented signals like intonation.

4.4.1 Cue Words Across Corpora

The corpus studies in the previous chapter and the perception experiments above indicate some broad distinctions in how different cue words are used. At opposite ends of the spectrum, affirmatives are used to admit propositions (or informational units) into the common ground, while negatives are used to reject them. In between these, non-affirmative *really* puts a (often weak) hold on such an addition, while *well* seems to signal that an adjustment is necessary for the dialogue to proceed. We can get another view into the meaning of these utterances by looking at their distributions with respect to different types of dialogue.

Table 4.4 shows the counts for the six cue words examined in the previous experiment from five different dialogue corpora. The Columbia Games Corpus (Gravano, 2009) and the HCRC Map Task corpus (Anderson et al., 1991) are highly task-oriented. Participants of these dialogues were given specific roles through which they had to work collaboratively to complete specific tasks. For example, in the Cards

game of the Columbia Games Corpus, one participant (the describer) described cards to another (the searcher) whose job it was to find a similar card on their board. The game concluded when all cards had been described. In the map task, one participant (the follower) reproduced a map route described by the other participant (the instruction giver). As with the Cards game, the participants did not have visual contact. In contrast, the participants of Switchboard dialogues (Godfrey et al., 1992) were engaged in rather more free conversation (see discussion in Section 3.2.1). Although they were given a topic to talk about, they did not have to stick to it and there was no designated completion point for the conversation in terms of achievements. The ICSI meeting corpus (Janin et al., 2003) includes multiparty conversations from meetings of various teams working on real ongoing projects. As such, these multilogues are more focused than Switchboard conversations but more open-ended than the games and map tasks. The sample from the Let's Go corpus includes telephone calls to an automated bus information service (Raux et al., 2005). So, these dialogues are task-oriented but also less collaborative than the other task-oriented sets.

There are two clear differences between conversational and task-oriented corpora. First, *okay* is the most frequent affirmative in the task-oriented dialogues, while *yeah* is by far the most frequent in the more conversational speech. Second, the conversational dialogues have a higher incidence of the non-affirmatives *really* and *well*.[5] These distributional differences between corpora are not so surprising when we consider what *really* and *okay* actually do. *Okay* is an affirmative for actions. This becomes apparent when we consider the types of moves that it can respond to. In the following we see that *okay* is decidedly odd after an information seeking polar question, but fine when the question is interpreted as an action request. In fact, it seems to be a preferred affirmative for imperatives.

(2) A: Did you do the laundry?
B: Yeah / # Right / # Okay / ?Sure

(3) A: Can you do the laundry before tomorrow?
B: ?Yeah / # Right / Okay / Sure

(4) A: Do the laundry!
B: ?Yeah / ?Right / Okay / Sure

So it makes sense that there are a higher number of *okay*s in the games and map task corpora because a large component of the dialogues involves one participant instructing the other. The more conversational speech captured in the Switchboard

[5]Only affirmative frequencies have been reported for the Columbia Games Corpus.

corpus is less instructive, so we get lower frequency for *okay* and a higher frequency of the positive proposition accepter *yeah*.

We can characterize the task in the conversational speech of the Switchboard variety as getting to know the other person. In this case, the objective is not so much to make sure all participants leave with the same beliefs, but instead to learn what the beliefs of the other participants actually are. So, learning something new and/or surprising is often a good thing. This seems to be why *really*s appear in this type of dialogue. In Section 3.5, I argued that *really* was underlying an interrogative. However, the information a *really* question seeks is already in the conversational background (if not the addressee's public beliefs), so its conventionalized use is to express that the information it is responding to is noteworthy in the sense of being unexpected. In this way it signals that the *really*-addressee is fulfilling the Gricean Maxim of Quantity in terms of being informative. Participants having divergent beliefs is generally well tolerated in the type of conversations found in Switchboard.

In the task-oriented dialogues, making sure that each participant's view of the common ground is crucial. If they diverge too much, the task will be difficult to complete. In the map task, for example, when the instruction giver says something surprising from the follower's point of view, it usually means that the follower is being unable to carry out the instruction. If a participant detects a break in common ground, then they need to actively fix it.[6] This usually requires a more explicit clarification request than what is produced by a *really*. Similarly, the check nature of *really* means that we don't expect it to see it used in the same backchannel way as it is in Switchboard. Hence we see a very low frequency for this cue word in the task-oriented dialogues. So, it seems we can use the frequencies of these sorts of cue words as indicative of the type of dialogue that is being carried out.

Beyond this it seems that different affirmatives underlyingly associate with different structures. Roughly following Portner (2007), we can postulate that like the usual sentence types, different cue words act on different components of the discourse structure. Declaratives, interrogatives, and imperatives present additions to the common ground, QUD and the to-do list respectively, while the cue word responses act on these proposed additions. We can characterize the differences between dialogue types in terms of structures: task oriented dialogues focus on the to-do list (i.e. actions), conversational dialogue focuses on the common ground (i.e. beliefs). *Okay* and *sure* act mainly as acceptance for the to-do list, while *yeah* and *right* accepts additions to common ground. *Really* and *well* can be used to bring up checks and modification on both of these structures. Similarly, we use the QUD to set the discourse topic, i.e. what participants are talking about at any point in the dialogue. Lack of resolution of the QUD may then reflect incomplete tasks on the to-do list or the fact that what's

[6]This is more or less inevitable in HCRC and IViE map tasks, because the maps don't match.

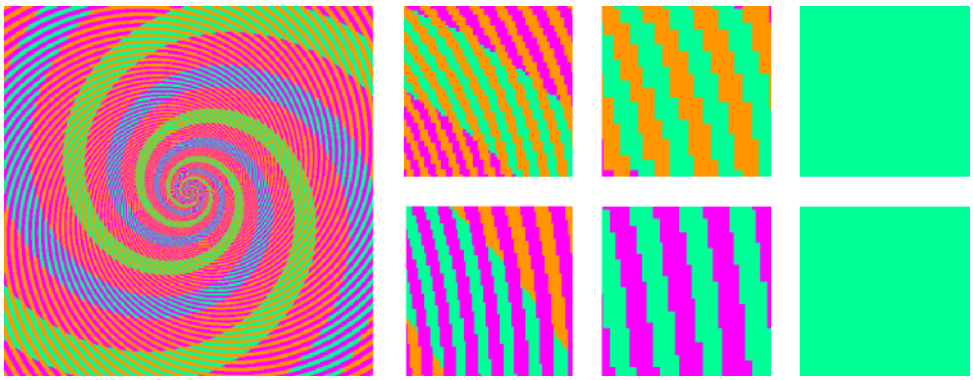


Figure 4.8: An Optical Illusion

already in the common ground is not sufficient to support a specific proposition under discussion.

4.4.2 Cue Words, Gradability and Evidence

We can derive a more precise notion of what these cue words mean by looking at their connection to how gradability is represented linguistically. This will help us get a grip on why some cue words are more amenable to prosodic influence than others. In determining the meanings of these cue words, the obvious place to start is in their sentential uses. In fact, adverbial *really* has already been analyzed in terms of its relationship with common ground.

***Really* and Standards**

Romero and Han (2004) separate epistemic, actuality and intensifier uses of *really* (the cue word falls in the former class). Examples of these different interpretations are given in the following (to be evaluated with respect to Figure 4.8).

(5) a. The blue and green spirals really ARE the same colour. (epistemic)
b. The blue and green spirals ARE really the same colour. (actuality)

(6) The blue and green spirals are really garish. (intensifier)

(7) A: The blue and green spirals are the same colour.
B: Really? (cue word)

Out of these interpretations, they exclude two major uses of *really* from their analysis: the intensifier (6) and actuality (5b) readings. They equate epistemic *really*

with the VERUM operator (which is also introduced by polarity focus and negative polar questions). They define VERUM(p) as a meta-conversational operator that says that the speaker is sure that p should be in the common ground. This is glossed as FOR-SURE-CG. This would seem to make *really* much like an affirmative.

(8) $[\![\text{VERUM}]\!]^{gx/i} = [\![\text{really}]\!]^{gx/i}$
$= \lambda p_{<s,t>}.\lambda w.\forall w' \in \text{Epi}_x(w)[\forall w'' \in Conv_x(w'')[p \in \text{CG}_{w''}]]$
$=$ FOR-SURE-CG$_x$

(9) The blue and green spirals ARE the same colour (verum focus)

While the intensifier version does seem somewhat removed from the epistemic version, the separation of the actuality case is more subtle. The actuality reading seems to arise when *really* appears after the focused auxiliary (negation, modals, or in negative polar questions), and expresses more that things are not what they seem, rather than speaker certainty about additions to the common ground.

(10) a. Mary isn't really human. (actuality)
$\not\approx$ The speaker isn't certain that Mary is human
$\approx$ Although Mary appears as such, the speaker knows that Mary is not human.
b. Isn't Mary really an alien?
c. Mary might really be an alien.

Like the intensifier, both epistemic and actuality interpretations seem to be involved in the setting of a standard, e.g. speaker certainty of the proposition under its scope. In this sense, *really* can take scope above and below negation (11). That is, in these cases the standard appears to be at issue.

(11) a. Mary really isn't a liar. (Tests confirm she is a human.)
b. Mary isn't really a liar. (She never said she was human.) $\neq$ Mary ISN'T a liar.

In (11a) *really* is used to affirm that Mary surpasses the standard for not being a liar, while (11b) expresses that Mary doesn't meet a standard for being a liar (not counting lies of omission). So, *really* seems more involved in setting the current standard of evidence used in discourse rather than asserting that a proposition belongs in the common ground.

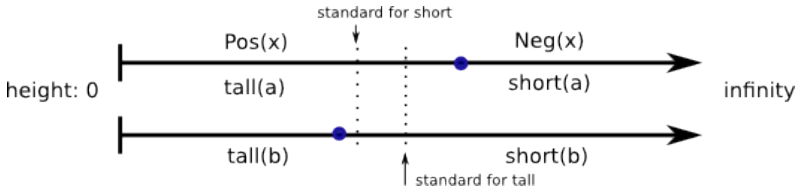


Figure 4.9: Evaluation of tall and short (Kennedy and McNally, 2005). Blue dots indicate the degree measurement for a and b. Here, a is tall and but b is short. Theoretically another object c could be neither tall nor short if it's degree falls in the interval between the tall and short standards.

Degrees and Propositional Probabilities

As an adjective intensifier, *really* is directly related to the notion of gradability. However, the adverbial version also seems to perform a type of intensification over beliefs.

(12) I really believe that chocolate is healthy.

The bottom line is that, *really* seems to say something about standards in all of these interpretations. We can use the approach of laid out in Kennedy (1999) and Kennedy and McNally (2005) to analyze the semantics of *really*. Kennedy (1999) argues that adjectives denote functions which map their arguments to some abstract representation of measurement, i.e. *degrees*. So, for some adjective ϕ, 'x is ϕ' means that the degree to which x is ϕ is at least as great as some degree $d_{s(\phi)}$, the standard for ϕ. This standard is determined by context. Kennedy argues that (i) degrees are intervals, and (ii) there are positive and negative degrees. This approach is visualized in Figure 4.9 for *tall* (positive degrees) and *short* (negative degrees).

Following this theory, unmodified APs contain a null degree morpheme, *pos*, from the the Degree Phrase, which relates the degree argument to a standard:

(13) **standard**$(d)(G)(\mathbf{C})$: degree d meets the standard of comparison for an adjective G with respect to a comparison class determined by $\mathbf{C}$, a variable over properties of individuals whose value is contextually determined.

(14) $[\![\text{pos}]\!] = \lambda G \lambda x. \exists d[\textbf{standard}(d)(G)(\mathbf{C}) \wedge G(d)(x)]$

This says that x has degree d that meets the standard induced by adjective G in the context of evaluation. Unlike other intensifiers like *very*, *really* seems to take any sort of gradable predicate that is not absolute. This might seem to rule out the use of *really* over propositions, since the degree that we would usually associate with propositions are truth values. However, with a bit more scrutiny it becomes apparent

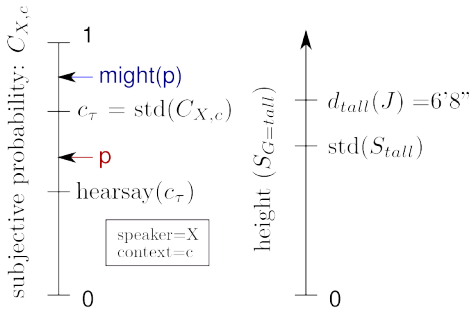


Figure 4.10: The quality thresholds of Davis et al. (2007) look like scales for gradable adjectives.

that even when we assert something, it doesn't mean that we have 100% certainty of its truth. Instead, it seems that we need to a certain level of confidence about it really being the case. The Gricean Quality required for admittance into the common ground can vary depending the task at hand. Davis et al. (2007) capture this with the notion of a *Quality Threshold*. This is a threshold on a subjective probability scale:[7]

(15) To assert p, X's subjective probability of p with respect to a conversational background c, $\mathbf{P}_{X,c}(p)$, should be greater than the Quality Threshold, c_τ, where

$$\mathbf{P}_{X,c}(p) = \frac{|p \cap \text{DOX}_{X,c}|}{\text{DOX}_{X,c}}$$

This is just the proportion of worlds in the speaker's doxastic worlds, i.e. the possible worlds that fit their beliefs, that are also p worlds. This looks a lot like the treatment of gradable adjectives in Kennedy and McNally (2005) (Figure 4.10). Davis et al. (2007) analyze evidentials as raising or lowering the *quality threshold*. Similarly, they suggest that epistemic modals MIGHT(p) and MUST(p) generally have higher probabilities than their non-modal counterparts, leading to their associated weak interpretation. In this vein, we might want to say that *really* signals that for proposition p, $\mathbf{P}_{X,c}(p)$ is high compared to other propositions deemed 'true enough'. However, the data suggests that there is a further evidential component to *really* that also needs to be accounted for. The crucial point is that sometimes you can't trust

[7]Recall that we follow the usual line in assuming that a proposition represents a set of possible worlds: the worlds where that proposition is true.

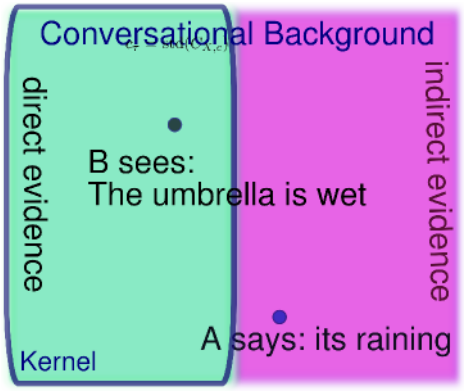


Figure 4.11: von Fintel and Gillies (2010) style kernel for epistemic modals.

the evidence most easily to hand. *Really* is used to manage the evidence that is used for evaluation of potential entrants to the common ground.

A baseline assumption would be that utterances are evaluated with respect to the best evidence the participants have. For example, we would expect direct evidence to trump hearsay evidence and atomic clocks to beat grandfather clocks in terms of timing. That is, we have some sort of ranking over evidence types. When it comes to evaluating propositions, we can cast this notion in terms of existing analyses for modals. In particular, von Fintel and Gillies (2010) argue that epistemic modals signal inference from the set of propositions representing direct evidence: the *kernel*. So, the argument is that there is an evidential component to epistemic modals which determines which propositions are selected from the greater knowledge background. In classic Kratzerian terms, the kernel is a form of modal base, i.e. a function from worlds to propositions that describes a conversational background for evaluation.

(16) $K(w) = \{p | p(w)$ and p is direct evidence wrt the speaker$\}$.

Then, MUST(p) indicates that the kernel entails p, i.e. $K(w) \subset p$. (c.f. Figure 4.11). von Fintel and Gillies further claim that epistemic modals carry a presupposition that the kernel does not directly settle the proffered content, explaining the infelicity of the following.

(17) (Alice is standing in the rain when she says)
??It must be raining.

Now, the line between direct and indirect evidence is somewhat fuzzy. For example, a speaker may be more likely to believe p if it asserted by a trusted source even if

direct sensory evidence suggests something to the contrary (cf. Figure 4.8). In this vein, *really* acts like epistemic modals in that they signal what sort of evidence is being used for the evaluation: it signals that the current standard of evidence has increased.

Really Raising the Standard of Evidence

We can then describe the sentential meaning of *really* as saying something about a set of possible worlds. More specifically, *really* raises the standard for what evidence can be used in an evaluation. We then generalize Davis et al.'s (2007) notion of propositional probability. Instead of using the full set of belief worlds, we use a contextually defined *kernel*, $f(w)$, as in von Fintel and Gillies (2010). The probability of a proposition is then:

(18) $\mathbf{P}(p \mid f) = \frac{|\cap f(w) \cap P|}{\cap f(w)}$.

Under von Fintel and Gillies (2010), the kernel associated with epistemic modals picks out direct evidence. Looking at the problem this way, we can easily transition from the categorical truth value evaluation of modals in that work to a probabilistic/gradable one: MIGHT(p) requests the probability that at least one world out of N draws is a p world, while MUST(p) requests the probability for drawing N p worlds.

For the problem at hand, the crucial point is that we can select other types of evidence. We take *really* to signal that evaluation is to be done with respect to better standard evidence than required by the contextual norm. That is, we have $f_{\text{really}}(w)$ which picks out a subset of the conversational background which would otherwise be used for the evaluation, i.e. $f_{\text{really}}(w) \subset f(w)$. By raising standards we drop evidence of a lower quality. This results in less restrictions which means more worlds in the evaluation domain since $\cap f_{\text{really}}(w) \subset \cap f(w)$. Expanding the evaluation domain reduces likelihood of 'accidentally' exceeding the standard. That is, having less assumptions means the argument of the *really* is more likely to truly have that property. This is visualized in Figure 4.12.

So, for gradable G which projects the degree of x to scale S_G, *really* acts like other intensifiers: it says the degree d exceeds a standard. For 'normal' gradable predicates, this means that the standard is created with respect to a larger comparison set. For a proposition p, the standard is the quality threshold, c_τ.

(19) $[\![\text{really}]\!](G)(x) = \exists d[\mathbf{standard}(d)(G_{\text{really}})(\mathbf{C}_{\text{really}}) \wedge G_{\text{really}}(d)(x)]$

(20) $[\![\text{really}]\!](\mathbf{P})(p) = \exists d[\mathbf{standard}(d)(\mathbf{P}(\cdot \mid f_{\text{really}}))(\mathbf{C}_{\text{really}}) \wedge \mathbf{P}(p \mid f_{\text{really}}) = d]$
$= \exists d[d > c_\tau \wedge \mathbf{P}(p \mid f_{\text{really}}) = d]$

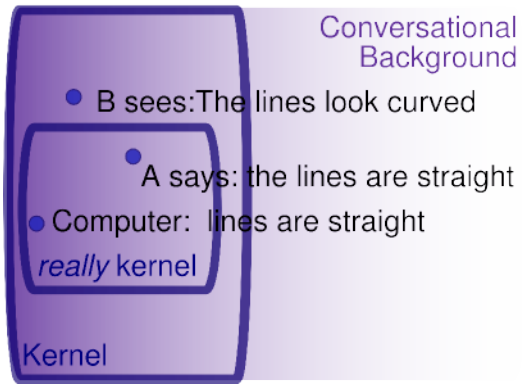


Figure 4.12: A generalized kernel for *really*.

(21) $\mathbf{P}(P \mid f_{\text{really}}) = \frac{|P \cap (\cap f_{\text{really}}(w))|}{(\cap f_{\text{really}}(w))}$

Recall that G is a function from objects to to degrees. When G is the usual sort of gradable predicate, we assume that the measurement is in a unit that doesn't change with context, e.g. meters. However, when it comes to conditional probabilities, the measurement does change depending on the discourse context, so the definitions above relativize the measure G to the kernel indicated by *really*. When the *measure* is absolute, the effect of the *really* is mostly in the selection of the comparison set. However, for the proposition version we assume that the quality threshold itself does not change as a result of *really*, but the domain of probability evaluation does. The following example shows an explicit change of domain for the adjective case.

(22) a. A: That tree is tall.
b. B: It isn't really tall. That giant sequoia we saw was tall.

Considering only trees in my garden results in a lower standard than considering trees in Philadelphia, which is again less general than considering all trees in the world.

A Note on Actuality

The relative height of *really* wrt modals sheds light on when/why you might get the actuality reading. The idea is that the position of *really* determines whether the evaluation is actually done with respect to the *really*-kernel for update purposes, or whether it just represents an alternative (possibly pedantic) viewpoint. So, the actuality case is about a possible standard rather than an established one. We can characterize the actuality reading in the following as saying that if we were to remove

some constraints (like things that look human are human) we might find out that Mary is alien.[8]

(23) a. Mary really might be an alien. (epistemic)
b. Mary might *really* be an alien. (actuality)

We can account for the fact that the difference in interpretation seems to correlate with height using the approach of Hacquard (2010) to the root/epistemic modal distinction. The basic idea is that modals take events as arguments. Modals above Aspect (epistemics) have their event variable bound by an assertion event, while modals below Aspect (root modals) are bound by the event generated by the verb. Similarly, *really* can appear above and below aspect, and when it is low the actuality reading is salient. So, *really* sets the evidential standard the event it is relativized to.

Really **as a Check Question**

In questions, *really* signals that what's already in the conversational background might deviate from reality. In the following example, a questioner might suspend direct auditory evidence.

(24) Did you really just say that?

This seems to arise when adding information to conversational background would mean a significant restructuring. Given what's in the questioner's (private) conversational background, p (or perhaps the act of uttering p) has a low probability. This is the sort of check that cue word *really* enacts. From here we can get a better idea of situations in which *really* has a backchannel sort of meaning. In order to interpret *really* as a backchannel, the *really*-hearer has to infer that their previous statements regarding p were sufficiently convincing. The probability of p in the *really*-speaker's previous model of the world is close enough to the Quality threshold - not much reorganization of beliefs has to go on. In this case, *really* just signals a small update to the conversational background. On the other hand, if the difference in likelihood is large, the *really*-speaker will need further evidence to justify the update. If this is sufficiently signalled, prosodically or otherwise, *really* will be treated like a real question. The following gives an example from Switchboard of *really* triggering this sort of post-hoc justification.

[8]Notice that *really* can act as a sort of intensifier on the modal. Again, this comes from the notion that expanding the set of worlds involved in the evaluation would make it harder to satisfy the conditions of *might*.

(25) A: What is your very favorite song that Billy Joel has done?
B: Probably Piano Man
A: Really?
B: Yeah
B: I uh m[aybe] maybe just because it is like on i think it is the first one on the on the CD

Similarly, *really*-speakers may follow up *really* with further questions. In the following, we see that more updating and more evidence is required than a simple yes or no answer.

(26) A: People think Charlotte is the big town you know and it's small
B: oh wow
B: really?
A: yeah
B: What's the population?
A: uh Charlotte?

The results the corpus and perception studies suggested that the questioning interpretation of *really* becomes more salient as its prosody becomes more effortful. This makes sense under the view that speakers put more effort into productions that are important for the dialogue to move forward as a *joint* task.[9] If a *really* is prosodically reduced *and* they don't think that the at issue content is controversial, the hearer is likely to interpret the *really* as a comment on the informativeness of the previous utterance rather than a real clarification request.

The Affirmatives

We can think of *right* in a similar sort of gradable way. As a modifier *right* offers a higher level of precision. In the following, *right* indicates that difference between the actual position of the ball from the line it was neglible.

(27) The ball was right on the line.

Following Kennedy and McNally (2010), we can characterize *right* as qualitative measure in that it gives us an indication of how far an object is from a limit point representing an ideal (meaning not far at all). In contrast, *really* can have a quantitative reading, in the sense of indicating a large degree rather than closeness to an

[9]Saying something under your breath may be satisfying, but it's not usually for the greater glory of the all participants!

ideal. From this point of view, *right* tightens standards in a more conventional way than *really*.

(28) a. That dress is really 80's.
(It has a lot of 80's features $\rightsquigarrow$ quantitative).
b. That dress is right from the 80's.
(It looks like it comes from that time period $\rightsquigarrow$ quantitative)

When *right* is used as a response, the affirmation indicates how accurate a characterization of the situation the previous utterance was. That is, the set of possible worlds picked out by this addition, fits the *right*-speaker's private model. Hence, the utterance not only passes the quality test but it actually reflects how that speaker sees the world to a good level of precision. This is what makes *right* a strong affirmative. It's not just used to accept propositions, it reveals something about the speaker. We can write this in terms of degrees as in (29).

(29) Right $\mapsto$ Evaluation is close to ideal.
a. $[\![\text{right}]\!](G)(x) = \exists d[\, G(d)(x) \wedge \mathbf{dist}(d, \text{limit}^C_G) < \epsilon\,]$
b. $[\![\text{right}]\!](\mathbf{P})(p) = \exists d[\, \mathbf{P}(p \,|\, f) = d \wedge \mathbf{dist}(d, 1) < \epsilon\,]$
i.e. $(1 - \mathbf{P}(p \,|\, f)) < \epsilon$, where ϵ is some small value for the relevant measure.

So, as a cue word response, *right* seems to need a measurement to comment on. That is, an evaluation of a proposition. This is something assertions provide, but true information seeking questions don't (cf. (30)). *Right* does improve, however, responding to a biased question. In response to an imperative, it gives the sense that fulfilling the instruction is a by-product of how things are. This can lead to a feeling of resignation.

(30) a. Do you dance?
b. Yes / #Right

(31) a. I'm not sure whether there's anything here for Sheila.
b. Doesn't she take dance classes?
c. Right, I forgot about that.

(32) a. Dance now!
b. Right $\rightsquigarrow$ I should dance now.

Unlike *right*, *yeah* doesn't seem to comment on measurement but acts as a more categorical accepter of positive polarity propositions, hence signalling that the proposition

meets the standard. It doesn't apply to units below the propositional level. *Okay* is more specific in that it highlights unexceptionality with respect to standards. The following examples show the differences between these affirmatives as modifiers.

(33) a. He's an okay/#right/#yeah dancer.
b. He's an ?okay/right/#yeah idiot.[10]

Okay rules out a 'good' evaluation. As a response the *okay*-speaker is accepting the other participants proposed action of adding a proposition to the CG. However, by using this particular response and not an affirmative like *right*, the speaker implicates that while they accept the move, they do not necessarily endorse the content.

(34) A: John dances so beautifully!
B: Okay.

Overall, we see that assertion type moves involve context dependent gradability. A proposition's evaluation can vary in given the speech situation, so that a statement that is acceptable in one context (e.g. small talk) is not acceptable in another (e.g. defusing a bomb). In conversation, speakers strive for the right amount of Quality for the job. Quality doesn't really come into play for the evaluation of imperatives, however, because the distinction between accepting and rejecting a task is sharper. Speakers may, of course, discuss future actions and add caveats, but in the end whether or not an instruction is carried out is up to the individual. It is plausible that *okay* is preferable in these sorts of situation because it doesn't add unnecessary overtones of Quality. However, even though we analyze *right* as dealing with degrees it is still appropriate as an affirmative in dialogue like the map task, like *yeah*, because accepting the state of affairs is similar enough to accepting a actions therein proposed.

Affirmatives and Rises

To make sense pragmatically in most cases, affirmatives like *right* need to affirm something from a source other than the speaker. Otherwise, we get the implicature that the speaker needed to put some extra effort to ground the information themselves. This is when adding a rise can actually help disambiguate the intention of the speaker. Putting a marker of non-finality onto an affirmative boosts the signal that the speaker needs help or ratification from someone else in adding their previous contribution to the common ground, and so we get a tag question interpretation. Without the rise, the default interpretation is that the speaker is doing this integration on their

[10]Constructions like *right idiot* are attested in in British English (OED).

own. In (35), Bert seems to be answering his own clarificational question with *right* (Bert is possibly being sarcastic). In (36), this sort of follow up is weird since the context suggests that the speaker is the source of the assertion and a clarificational accommodation is difficult.

(35) Alice and Bert are watching a show about street food in Cambodia.
A: You like this sort of thing.
B: I like deep fried spiders. Right.

(36) A: What type of snacks do you like?
B: ?I like deep fried spiders. Right.

In response form, the perception experiments suggested that *yeah* is more affected by rises than *right*. I previously claimed that this was because *yeah* has a less defined semantics than the other affirmatives. We can see now what this means. *Right* and *okay* are strong in that they give a fairly specific evaluation of the circumstances, while at core *yeah* reveals less about the speaker's attitude. A *yeah* response may indicate a strong agreement or a borderline acceptance. *Right* indicates that the speaker finds an utterance to be credible, but moreover that it is settled given the *right*-speaker's world view (since we take *right* to express a level of accuracy). So, the possibility that a rise would signal propositional uncertainty and instead rises seem to simply signal that the speaker expects more talk. That is, rise is points to a lack of resolution of a higher level question under discussion, rather than the current one associated with the at-issue content.

(37) A: Why did you go to that stall?
B: Well, you know I like spiders...
A: Right...
B: They have deep fried ones there.

For *yeah*, there is less blocking an interpretation pointing to openness with respect to the at-issue content so we get more of a check question interpretation.

4.5 Conclusion

The goal of the studies presented in this chapter was to shed light on how cue words and prosodic features affect dialogue structure and maintenance. The results of the perception experiments supported the idea that prosodically effortful features highlight the underlying meaning of the utterance, rather than adding a specific attitude.

The perception of surprise positively correlated with prosodic effort, as represented by pitch range, for *really*. Similar prosodic effort did not result in the same levels of surprise/questioning for *right* for which the ratings generally stayed on the bottom half of the scale. The high ratings for on both question/surprise scales *really* confirmed that idea that this cue word is underlyingly interrogative. In fact, surprise ratings were correlated with how questioning the *really* sounded, however this was not correlated with the final rise/fall characteristic.

The results of the second perception experiment supported the idea that rises signal some lack of QUD resolution, rather than specifically expressing uncertainty about the credibility of the content the cue word is directed at. Cue words themselves vary on the credibility scale: *right* indicates greater credibility than other affirmatives like *okay*. Looking at the interaction of cue words and contour type it appears that properties that have been associated with rises more reflect the properties of the underlying content and the discourse situation. Playing a check move, like *really*, for example, already shows some level of uncertainty about the quality of the previous utterance. So it seems that, rises/falls reveal the status of the QUD, while cue words signal the level of acceptance of respondent. Overall, the data support the idea that rises signal that the discourse has not come to a viable stopping point.

In general, understanding how these parts of dialogue compose has implications for both formal theories of dialogue and for determining how a dialogue system should respond to such cues of speaker attitude. The perceptual results shine a light on why specific cue words are used when they are. We got a clearer view of these differences by examining how these utterances related to different discourse structures. This was brought out by looking at frequencies of use in different dialogue types. In general, we saw that the less goal oriented the dialogues, the more tokens of *really* we found. At the other end of the spectrum, we saw more occurrence of *okay* in more task-oriented dialogues. This in turn seems to reflect the fact that *okay* is an affirmative for actions, while *yeah* is primarily an affirmative for propositions. Moreover, the different ways these responses are used bring out how evaluation of utterances in dialogue is a matter of degrees, rather than absolutes. In dialogue, truth is gradable.

Looking in detail at the meaning of *really* outside of the cue word use showed how we can we can bring this into the fold with other formal treatments of gradability. This give us a way to think about the meaning of these cue words in a more precise way, which in turn leads to a clearer way of thinking about their interaction with prosody. We also get a peek under the hood. This sort of analysis highlights what other structures we need to calculate meaning below the level of the common ground and to-do list. This highlighted the fact that across and within dialogues, standards are changeable. To really model a conversation we need to keep track of these standards, especially those that determine what sort of evidence is admissible at any given time,

and how precise our descriptions are expected to be.

Overall, the studies of cue words presented in this chapter and the previous have demonstrated that intonational forms, particularly terminal rises, act at the level of discourse structure. Moreover, we found attitudinal variables to be useful in investigating how variation in prosody relates to variation in meaning. However, that attitude derives from the underlying meaning of the utterance. Cue words were a good probe for these studies because they are used frequently in dialogues, and like prosody, their primary use appears to be discourse management. However, the flip side of this is that their discourse uses are by nature conventionalized. So, we would like to know whether our results on prosodic meaning generalize to situations where the utterance meaning is compositional. As such, in the next chapter we turn our attention to the prosody of sentential utterances.

Chapter 5

Declarative Responses and Prosody

5.1 Introduction

In this chapter, we will continue to probe the question of what and how terminal rises contribute to utterance meaning, turning our attention to full declarative responses. So far we have been looking at dialogue moves that deal wholesale with propositional units of information. However in order to carry out these moves we must also look at subsentential informational units. That is, we need to look at the Information Structure (IS) of the utterance and its relationship to prosodic form. Unsurprisingly, prosody plays an important roles in theories of Information Structure, particularly with respect to how notions like topic and focus relate to pitch accent and boundary tone varieties (Roberts, 1996, Steedman, 2000, Büring, 2003, a.o.). As was the case for terminal rises, a crucial problem brought up by theoretical IS and discourse work is how tightly we should bind specific intonational forms to theoretical linguistic constructs. Results from the previous chapters suggest that the answer is not very tightly. However, we would like to see if those results extend to situations where the underlying semantics of the utterance is not so conventionalized as it is in the cue word case, but instead where the semantics is properly compositional. Moreover, other experimental studies indicate that IS does play a consistent role in determining the prosodic form of an utterance. For example, based on experimental data, Calhoun (2007) argues that the topic/focus difference can be characterized in terms of relative prominence instead of accent/boundary shape.

So, to understand the contribution of sentential prosody in this sorts of utterances we will also have keep a grip on their IS properties, and also the relationship between

IS and pitch accents. To keep this investigation in line with the work in the previous chapters, we will concentrate on an IS configuration that is as close as possible to the discourse/propositional level that we treated cue words. Such a configuration appears in the form of Verum Focus (VF). Verum focus broadly covers utterances which seem to emphasize the truth of a proposition. Several semantic/pragmatic analyses have cast VF as a special sort of discourse operator, requiring a different treatment to 'normal' types of focus. In fact, as mentioned in Section 4.4.2, Romero and Han (2004) equate VF with epistemic *really*. I will argue that once we take into account the broader discourse requirements *and the structuring role of prosody*, it makes more sense to treat VF as we would any other type of information structural focus, where the focused element is propositional polarity.

With this in mind, we would like to know how both higher level discourse structure and IS requirements change the expectations of what intonational forms are available. Impressionistically, declarative responses like (1b) and (2b) are produced with different contours even though they have identical propositional content and perform the same speech/dialogue act (assertion/statement/inform).

(1) a. So, Emily brought a meringue.
b. Right. Emily did bring a meringue

(2) a. Nobody brought a dessert.
b. Emily did bring a meringue

(3) a. Emily brought a meringue.
b. What? Emily did bring a meringue?

The intuition is that the *direct agreement* in (1b) is produced with falling pitch from 'did', while the *indirect contradiction* response (2b) naturally has an extra fall-rise accent on 'meringue'. Furthermore, using the intonational contour of (1b) sounds infelicitous in the context of (2). Both of these differ from the declarative question (check move) in (3b), which is intuitively produced with a rising accent. So, it seems that these sorts of response dimensions may be more consistent predictors of meaningful prosodic variation than the usual type of dialogue act. Moreover, I claim that folding VF into the more general treatment of IS focus this way sheds light on why we get the prosodic facts about it that we do. However, to support this idea we need to know how robust these intuitions are. Furthermore, we would like to know whether the intonational form is affected by whether the response is an agreement or contradiction.

These issues are investigated via production and perception experiments through this chapter and the next. The goal is to find discourse configurations that consis-

tently associate with rises. We also want to know what sort of prosodic configurations produce the same expressions of attitude (discourse or propositional) in a given context. Looking at the phonetic detail, we find that fairly distinct nuclear tune patterns associate with different response types. However, as with the cue words, we find that the presence of a rise itself is not enough to distinguish a particular type of discourse move or IS unit. Instead the data supports our hypothesis from the previous chapters that terminal rises primarily signal non-finality. This, again, highlights the role of discourse structure in determining move type. The studies have implications for the assignment of meaning to the intonational manifestation of utterances. In particular, the productions of indirect responses bear on the proper analysis of the fall-rise (FR) accent (Jackendoff, 1974, Ward and Hirschberg, 1985, Büring, 2003, Constant, 2007, a.o.).

This chapter is organized as follows. Section 5.2 reviews relevant notions about information structure, metrical structure and prosody. Section 5.3 discusses verum focus as an IS configuration that like cue words, says something 'about' information at the propositional level. We will look at how the different types of responses it is involved in relate to specific intonational forms, and how this relates to its discourse/IS properties. Section 5.4 presents a production experiment investigating how well these intuitions about different verum response types match real data. As in the previous corpus studies, we look at the variation in the data and the separability of the different discourse categories. With respect to the latter, we use classifiers to examine the robustness of these differences and to explore the variation in the data. This allows us to test hypotheses about the whether specific intonational shapes map to specific discourse categories. Overall, the data argues against any one-to-one mapping between abstract intonational forms and categories like topic/focus or agreement/contradiction. However, we do find that prosody varied consistently depending on response dimensions like whether an utterance directly or indirectly addresses the question under discussion. These findings are discussed and summarized in Section 5.5.

5.2 Information Structure and Sentential Prosody

When dealing with a cue word like *really*, we more or less know where a pitch accent will fall (i.e. the first syllable). Given that we know where it is going to fall, we can characterize the intonational form of the utterance in terms of the F0 height, tilt and convexity around that prominent position. As we turn our attention to larger, sentential utterances we have work a little harder to do this characterization. Sentential utterances may have more than one pitch accent, and the placement of these

prominences seems to depend on a mix of sentence internal and external factors. A key pointer to this is Question Answer Congruence (QAC), illustrated in the following. The type of wh-question matches the placement of stress in the answer. In (7) the placement of stress determines what item is under clarification.

(4) a. Who stole the money?
b. JOHN stole the money

(5) a. What did John steal?
b. John stole the MONEY

(6) a. Who stole what?
b. JOHN stole the MONEY

(7) a. JOHN stole the money?
b. John STOLE the money?

The basic idea is that what receives the main sentence stress in an answer depends on what question the sentence is answering. A good deal of work on information structure and prosody from the theoretical point of view, has gone into predicting where and when these prominences will fall. To talk about these things we need to define the basic information structural categories we are dealing with: topic and focus. Unfortunately, the abundant literature on this topic contains an equally abundant collection of definitions for these terms. In the following we will follow the basic approach of Vallduví (1990) (cf. Steedman (2000), Eilam (2011)) for describing the information components of a sentence. In Vallduví's framework, IS represents what updates to participants' knowledge stores are to be carried out. That is, how participants track informational objects below the propositional level. The knowledge store can be thought as a set of filecards (Heim, 1983), where each card contains information about a specific entity in the conversational background. We can roughly take the following declarative 'John stole the money' as adding the information that John stole the money to the filecard about John.

More generally, we can partition an utterance into focus and ground, where the ground can be further separated into topic and tail:

(8) **Focus:** The information to be added to the file card, i.e. the update/answer to the current question.

(9) **Ground:** Indicates where the update is to be done in the knowledge store.
a. **Topic:** The filecard address, i.e. what the utterance is about.[1]

[1]This was originally called the *link* in Vallduví (1990). I take Eilam's (2011) naming here since

b. **Tail:** Further information on how the update is to be done, i.e. the leftovers.

Under this view, *the money* is the focus of B's utterance, and the rest makes up the ground. Out of this, *John* is the topic, and the tail tells us that the focus should be added to the denotation of things stolen by John.

(10) A: What did John steal?
B: $[\text{John}]_{\text{Topic}}$ stole $[\text{the MONEY}]_{\text{Focus}}$

The notion of focus given above is an information packaging one and is closely aligned with the notions of rheme (= focus) and theme (= ground) of Steedman (2000). However, focus also appears frequently defined in terms of alternative semantics (Rooth, 1985). This sort of analysis arose from attempt explain the connection between prominence and so-call focus sensitive operators, e.g. *only*, *even*, *also*. The following examples show how the placement of prominence can change the truth conditions of the utterance.

(11) a. John only *stole* the money.
(he didn't burn the money, he stole a car)
b. John only only stole the *money*
(he didn't steal the car, he burnt the money)

These operators are assumed to quantify over sets of alternatives. To account for this Rooth (1985), following Jackendoff (1974), argued that focus prominence determined the alternative sets available to the operator. The offshoot of this is that focus is taken to generate alternatives in a special semantic representation of the utterance: the focus semantic value.

(12) John only stole the MONEY
a. Ordinary semantic value = $\{w \in W \mid \text{John stole the money in w}\}$,
b. Focus semantic value = $\{\{w \in W \mid \text{John stole x in w}\} \mid x \in E\}$,
where W = the set of possible worlds, and E = {the car, the jewels,...}

In light of the QAC, it makes sense that IS focus evokes alternatives (i.e. other possible answers to the question), however we can see that this type of alternative generating unit doesn't always match up with the IS focus. In the following, the IS focus is the place requested by the *where*-question, Tokyo. Putting the main sentence prominence on the associate of *only* instead, and so breaking QAC, is infelicitous.

it is more transparent.

(13) Where are there only skyscrapers? (Eilam, 2011, (56))
   a. There are only skyscrapers in TOKYO
   b. #There are only SKYSCRAPERS in Tokyo
   c. In TOKYO, there are only skyscrapers.

Again following Vallduví (1990), we treat this sort of alternative generation as a separate IS dimension: background versus kontrast (F-marking in Calhoun (2010), semantic focus in Eilam (2011)).

(14) **Kontrast:** Kontrastive (or **F-marked**) elements generate alternatives to the marked element at some level of semantic representation.

(15) **Background:** The parts of the utterance which do not participate in alternative generation.

Note: In the following, we use subscript F to label F-marked elements, i.e alternative generators, and CT for Büring style contrastive topics. Subscript *Focus* or ρ (rheme) *Topic* or θ (theme) mark the IS focus and topic respectively.

5.2.1 Pitch Accents and Metrical Structure

So it seems that the notions of IS focus, F-marking, and pitch accenting are deeply entwined in English. What exactly their relationship is has been a question of much debate. A baseline analysis is that pitch accents are a direct manifestation of some structural feature, usually given birth to and percolated around in the syntax. It is not hard to construct examples showing that not all pitch accents mark IS focus.

(16) a. What happened?
   b. Your friend from FRANCE brought us a bottle of WINE.

So our next guess would be that such a feature should represent F-marking. This is the line of attack led by Selkirk (1986). In these sorts of accounts pitch accent marking is taken to be independent of metrical structure. Calhoun (2007) presents evidence against this approach, arguing instead for a metrical-structure based approach (Liberman and Prince, 1977). I will refer the reader to that work for the detailed argument, but we can briefly outline some of reasons why a metrical approach is preferable. The main benefit of this approach is that it can easily deal with cases where prominence doesn't seem to lead to alternative generation. In (18), 'mother in law' naturally acquires prominence both when it is in a contrastive relationship (F-marked) and when it is not.

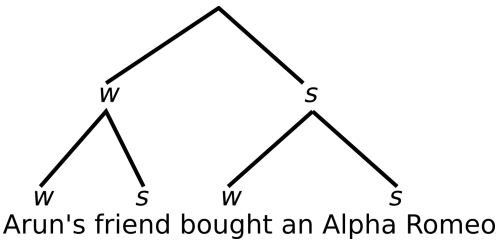


Figure 5.1: Right branching metrical structure (broad focus) following Calhoun (2007).

(17) [Q:] What did Arun's mother-in-law think?
A: (Arun's MOTHER-in-law [DISAPPROVED]$_F$)
A: ([Arun's MOTHER-in-law]$_F$ [DISAPPROVED]$_F$
(but his [FATHER-in-law]$_F$ [LOVED it]$_F$)

Cases where a prenuclear pitch accent is contrastive are difficult for the non-metrical accounts. Clearly some prenuclear pitch accents are F-marked and some are not, however there is no principled way to identify which are which. However, they are predicted in an account where prominence is assigned *relatively*, based on a hierarchical structure. In this way, we predict relative metrical strength by the number of strong nodes above a given leaf in the tree. The assumption is that English metrical structure is right branching by default (cf. Figure 5.1). This gives a principled reason for why in broad focus contexts – where the answer to the question is the whole proposition – the default place for the accent is at the utterance end, however a large enough metrical structure may induce pre-nuclear accents.

In Calhoun's analysis, the discourse properties of words in the utterance affect how that string is mapped onto metrical structure, so foci usually, but not always, map to nuclear accents. Deviations from the expected structure result in contrastive effects. For example, a prominence on 'friend' in the following broad focus response doesn't seem to lead to a contrastive reading, but prominence on 'Arun' does.

(18) a. What happened?
b. (Arun's friend bought an Alpha Romeo)$_F$

To summarize, we take the following basic view of the relationship between information structure and pitch accent placement.

(19) a. There are two basic IS dimensions:
(i) IS focus/ground,
(ii) background/kontrast=F-marking.

b. Pitch accent placement is guided by the metrical structure.
c. Discourse properties of the utterance affect the metrical structure that the words in that utterance obtain.
d. Deviations from metrical expectations induce contrastive (F-marked) interpretations.
e. The IS focus (usually) maps to the metrically strongest position.

The IS notions outlined above deal with organization of information and prosody with a single move. Given that our our working hypothesis is that terminal rises act on discourse structure above that level, we would like to know how different move types affect how IS is manifested. In particular, we want to know how this relates to the interpretation of terminal rises. Cue words like *really* and *right* comment on propositions as a whole. These one word responses also seem to be quite conventionalized in the type of move they play. *Really* basically plays a check move, while *right* expresses agreement. For sentential responses, more options are opened up because the relationship between an utterance and what it is responding to can be more complex. For example, a single sentential utterance can be interpreted as an agreement or a contradiction depending on the surrounding context. Cue words provide direct evaluations, while sentential utterances may be more indirect in how they address the QUD. To keep things comparable with the cue words, we want to look at the relationship between prosody and different response types while keeping focus on propositional level informational content. As such, the next section examines these issues by investigating the meaning and discourse properties of utterances bearing verum focus.

5.3 Information about propositions: Verum Focus

Verum Focus (VF) is a broad term which covers utterances which seem to emphasize the truth value of the proffered content of that utterance. In English this usually manifests as emphasis on the copula or with *do* insertion. That is, emphasis on the carrier of tense.

(20) Mary IS an alien.
⇝ It's true that Mary lost her memory.

(21) Mary DID lose her memory
⇝ It's true that Mary lost her memory.

These sorts of utterances can be used in both direct affirmations (22) or contradictions (23).

(22) A: So, Marianne met with Lenny (Affirmation)
B: Right. Marianne DID meet with Lenny.

(23) A: Marianne didn't meet with Lenny (Contradiction)
B: No! Marianne DID meet with Lenny. Look at the logs!

Like a cue word *right*, verum focus then seems to explicitly mark the proffered content as being in the speaker's public commitments. As such, it forms a natural class with respect to information packaging: verum focus utterances are *about* the truth or falsity of propositions. This translates to being about whether or not a proposition should be accepted in the discourse. From an IS viewpoint VF sets the ground to be the proffered proposition while the IS focus is the truth value itself. So, like our cue words, VF utterance seem to make a comment about a proposition that is already in play.

Verum focus occurs naturally in situations where there has been an update in the polarity due to new evidence, e.g. (24), or where the speaker has a contrary opinion to the hearer, e.g. (25). In the former case the speaker can be seen to be publicly updating their own beliefs about the polarity of the proposition, while in the latter the speakers attempts to update the hearers beliefs. In both cases the speaker attempts to add a proposition to the common ground. Note, we do not get the same effect if we simply focus the finite verb, e.g. (24b).

(24) Alice and Bert are speculating on the after effects of Mary's recent accident. It seems she may have experienced some memory loss, but they aren't sure. They now see that she doesn't recognize any of her relatives
a. A: She DID lose her memory.
b. A: #She LOST her memory.

(25) A: Mary's just faking the amnesia.
B: No, she DID lose her memory.

So, verum focus appears to mark the truth value of a proposition as a previous object of discussion. As such, utterances containing verum focus generally seem incongruous out of the blue. However, example (24) shows that they are are licit when the context provides grounds for a truth value update even without immediately previous talk. In such cases, however, the use of verum focus still suggests that the proposition has already been under consideration by (at least) the speaker. So, it makes sense to

consider the meaning and use of verum focus declaratives in terms of what sort of questions they can answer. This is something we need to sort out in order to get a grip on how different prosodic forms affect the interpretation of these sorts of declaratives. In particular, this will help us get an idea of what sort of contexts induce terminal rises.

5.3.1 Verum Focus Responses

Direct Responses

Given that the examples above require that the VF proposition already be under discussion, it is not too surprising that VF can produce infelicity in answers to wh-questions. In (26b), for example, there does not seem to be any reason for Bert to highlight the polarity of his answer given that the content of the answer itself, 'John finished the book', is new to the discourse. However, verum focus is fine if the answer content is already under discussion, as in (27). That is, the VF utterance is both an answer to the wh-question and a salient question under discussion.

(26) A: What did Carl do on the weekend?
   a. B: He finished the book.
   b. B: # He DID finish the book.

(27) Alice is very clearly anxious about Carl, a notorious procrastinator, finishing his book before today's deadline. At the same time she doesn't want people to think she cares, though everyone knows that she does. Alice sidles up to Carl's roommate, Bert.
   a. A: So Bert...What did Carl do on the weekend?
   b. B: He DID finish the book, since you're wondering.

Unlike wh-questions, polar questions explicitly ask for a polarity judgement: a polar question $?p$ asks the addressee to select between $p, \neg p$. So, we would expect verum focus to be licit in answers to polar questions and this is the case, e.g. (28).

(28) a. A: Did you write the review?
   b. B: I DID write the review.

We can treat direct VF denials and agreements as basically being the same as VF answers to polar questions (i.e. a hearer either accepts p or rejects p). This suggests that the main constraint on VF(p) is that p already be under discussion, and so in these scenarios a speaker is attempting to *change* the polarity of p to true.

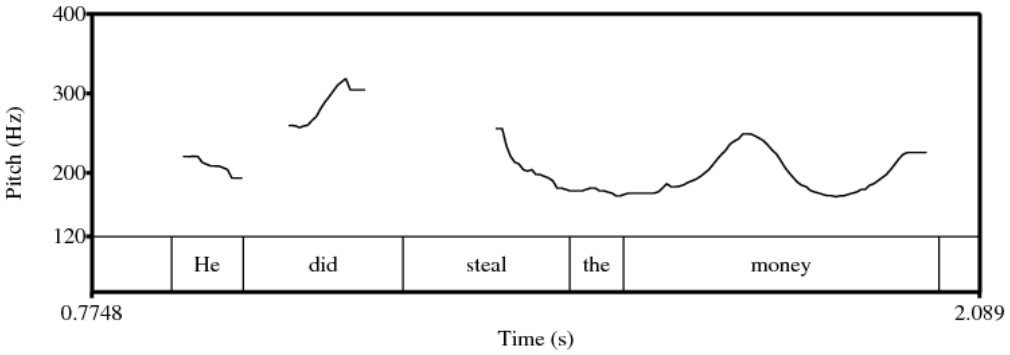


Figure 5.2: Fall-rise in example (29): 'He DID steal the *money...*'

Indirect Responses

Unlike the direct answers above, indirect affirmations/contradictions using verum focus do not seem to indicate a change in polarity. Instead, they bring up information that is relevant for evaluation the current question under discussion. For example, in (29) Bert's answer suggests that evaluation of John's honor needs to take into account the fact that he stole money (besides other facts such as his feeding the poor). However, polarity focus still seems to be used here to highlight the truth of the proposition.

(29) A: Is John an honorable guy?
B: He DID steal the *money...*
B: But, then he used it to feed the poor.

This sort of response seems to require a Fall-Rise (FR) accent, e.g. Figure 5.2.[2] Such FR utterances are infelicitous in polarity update situations. For example, an FR accent on *memory* in is decidedly odd given the context of (24), reproduced as (30) below.

(30) Alice and Bert are speculating on the after effects of Mary's recent accident. It seems she may have experienced some memory loss, but they aren't sure. They now see that she doesn't recognize any of her relatives

[2]Notation: I will use SMALLCAPS to indicate 'normal' falling/H* accents and *italics...* indicate a the domain of a fall-rise (Constant, 2007):

(i) a. He STOLE the money. (fall)
b. He *stole* the money... (fall-rise)

a. ??A: She DID lose her *memory*...

Note: This sort of verum declarative carries two pitch accents: a fall on *did* and a fall-rise on *memory*. I will refer to this type of contour as verum-FR. To be licit as a direct answer to a polar question, a verum-FR declarative needs to be interpreted as addressing a subquestion (cf. Section 2.3.2). That is, part of an answer to a dominating question, e.g. *What did you do?* in (31).

(31) A: Did you write the review?
a. ??B: I DID write the *review*...
b. B: I DID write the *review*... I just didn't prepare the handout.
⇝ Dominating question: What did you do (out of the relevant tasks)?

Similarly, in (32) verum FR is fine as an affirmative answer to an implicit subquestion ('Does John like poker') after a series of negatively answered related questions. Note, that the same utterance without the second prominence, e.g. (32b), does not sound appropriate.

(32) A: There must be some sport that John likes...
A: Does he like basketball?
B: No.
A: Football?
B: Not really.
A: Baseball?
B: No.
A: Does he like tennis?
B: No.
a. B: He DOES like *poker*...
b. B: ??He DOES like poker.

Propositional Contrast

VF also pops up when the polarity of one proposition is being directly contrasted to another. In the following, the VF utterance can also be seen as an indirect or partial answer to the immediate question under discussion. Note, this does not seem to require a terminal rise. If we add one, it seems to suggest, like the examples above, that the current QUD is still not resolved. Without it, we assume that the two clauses do resolve the current QUD, from the VF speaker's perspective. For example, B, in the following, may be unsure whether having done the vacuuming is sufficient to say that he has done his chores (having not done the laundry).

	Verum		Non-verum	
	word	count	word	count
1	and	138	and	146
2	but	125	i	93
3	i	112	but	60
4	well	46	so	39
5	so	34	well	35
6	you	17	that	34
7	we	16	we	30
8	it	15	it	28
9	uh	14	they	20
10	they	11	you	20
11	oh	10	uh	15
12	that	10	oh	12
13	um	10	the	11
14	the	7	he	4
15	if	6	um	4
16	or	6	if	3
17	now	5	now	3
18	he	3	she	3
19	there	3	there	3
20	when	3	this	3

Table 5.1: First word of the highest level S for verum, non-verum utterances (SWBD-NXT/Treebank)

(33) A: Did you do your chores?

a. B: I DIDN'T do the LAUNDRY, but I DID vacuum the LIVING room.

b. B': I DIDN'T do the LAUNDRY, but I DID vacuum the *living* room...

We see this contrastive use of VF reflected by a relatively high co-occurrence with the contrastive discourse marker *but* in the Switchboard data. Table 5.1 shows the twenty most frequent initial words for verum type *do*-insertion utterances in the Switchboard corpus (based on the highest level S node dominating the inserted *do*) and a sample of non-verum utterances from the same set of conversations (627 utterances, each). Inspection of the corpus data suggests that this indirect use of VF is the most prevalent in this style of speech. This reflects the fact that there is less of an explicit question/answer structure in conversational speech.

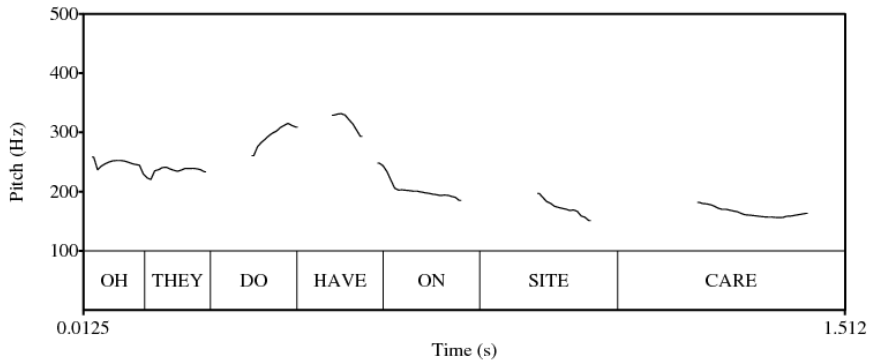


Figure 5.3: A classic verum update contour labelled as a question in SWBD-MDE (sw2943).

Declarative Questions

As we would expect, VF in declarative questions appears to check the polarity of a proposition. Again, use of VF seems to imply that the proposition was formerly under discussion, and evaluated as either unresolved or negative.

(34) Marianne DID do the dishes?

Like other declarative questions, the default nuclear accent is convex through the IS focus and rising to the end of the utterance. However, this intonational form is not necessary for the utterance to be perceived as a question. In the following example from Switchboard, speaker A seems to express surprise that B has workplace childcare. We can see this as A trying to set *B's husband's company has on site care* to True. The fact that she is doing this explicitly suggests that this valuation is unexpected. In fact, this was annotated as a question even though it seems to show the classic falling update contour (Figure 5.3).

(35) (sw2943)

B: and they have you know a deduction kind of plan where it's tax free you know if you put it in ahead of time

...

B: and on site care

A: uh-huh

B: those kinds of

A: oh they do have on site care (question)

B: no no

B: they don't

So we see that the context gives the VF-utterance the flavour of a declarative question rather than the intonational pattern: it's seems assertive in that A is trying to add something to the common ground, but it's about old information (since B seems to have said it).

Verum Focus and Narrow Focus

In summary, we have two basic types of contexts for verum focus answers. In the direct responses polarity is updated/changed. In these cases, only the inserted auxiliary is prominent. In indirect responses, the VF utterance does not seem to be about polarity update, instead it brings forward some relevant evidence from the conversational background about a dominating question under discussion. Here, multiple prominences are possible. Moreover, a terminal rise is possible while still retaining assertive status. Both of these cases are about contrast in polarity. In the direct responses, a newly acquired positive polarity is contrasted with an old value. In the indirect case, the contrast seems to be with respect to the polarity of sister subquestions.

Looking a bit more broadly, it seems that VF exhibits the general restrictions and implications of narrow focus. Narrow focus utterances are good for updates expressing contrast or surprise (36). Like the VF data, narrow focus is generally weird out of the blue (37), and similarly infelicitous when answering the 'wrong' question (38), recalling the congruent question for VF is one about polarity.

(36) Who would have thought!
  a. *Mary* did the dishes, not Alice.
  b. Mary *did* do the dishes.

(37) Hey Bert!
  a. ?*Mary* did the dishes.
  b. ?Mary *did* do the dishes.

(38) What did Mary do?
  a. ?*Mary* did the dishes.
  b. ?Mary *did* do the dishes.

This pattern seems to lead back to the notion of question/answer congruence, rather than any special properties of verum focus. So, it seems like we can treat VF as

focus on polarity. That is, for VF(p) the IS focus is the truth value of p while the topic is p itself. This means that non-tense associated prominences arise for reasons of rhythm or kontrast. This in turn depends on the current discourse structure and QUD. However, several analyses have argued that verum focus is a different sort of semantic operator to 'normal' focus. In the following we review these analyses and compare them to the informational structural account.

5.3.2 Previous analyses of VERUM

Romero and Han (2004) claim that various linguistic objects, e.g. polarity focus, preposed negation in Negative Polar Questions (NPQs), and *really*, introduce a semantic operator, VERUM (cf. Section 4.4.2). This operator is *metaconversational* in nature. That is, it relates to the conversational goals of the speaker. R&H give VERUM the following definition, glossed FOR-SURE-CG$_x$:

(39) $[\![\text{VERUM}_i]\!]^{g x/i} = [\![\text{really}]\!]^{g x/i}$
$= \lambda p.\lambda w.\forall w' \in \text{Epi}_x(w)[\forall w'' \in Conv_x(w'')[p \in \text{CG}_{w''}]]$
$= \text{FOR-SURE-CG}_x$.

By this definition VERUM(p) signals that p is in the common ground in all worlds where the conversational goals of speaker x are met. That is, the speaker is certain that p should be added to the common ground.[3]

The main motivation for this analysis of VERUM is to explain the bias and negative/positive polarity item (NPI/PPI) licensing facts associated with NPQs. NPQs are, in principle, ambiguous between negative and positive bias readings. Inner NPQs (INPQs) project an expectation of a *no* answer, while Outer NPQs (ONPQs) expect an *yes* answer (Ladd, 1981). These two readings are made more salient by the presence of NPIs (e.g. *either*) and PPIs (e.g. *too*), the former of which are only licit in INPQs.

(40) a. A: It was a pretty poor turn out at the practice talks. John and Mary didn't show up, so there weren't any phonologists.
b. B: Didn't Jane go either? (Inner NPQ, expect *no*)

[3]This is different from simply saying that the speaker is sure that p is true. The motivation for this is the difference in meaning between the following pair. However, note that this already assumes that VERUM and *really* are the same thing.

(i) a. I am sure I am tired.
b. I really am tired.

(41) a. A: John went to the talk but Mary didn't, so there weren't any phonologists.
b. B: Didn't Jane go too? (Outer NPQ, expect *yes*)

In R&H's approach, this ambiguity is characterized as a scope ambiguity between the VERUM operator and negation. In INPQs, VERUM scopes above negation and, since nothing is between negation and the clause, NPIs are licensed. This is not the case for ONPQs where VERUM intervenes. This results in the following denotations for the NPQs above, (40b) and (41b).

(42) a. Didn't Jane go either?
$\{\text{FOR-SURE-CG}_x(\neg\text{p}), \neg\text{FOR-SURE-CG}_x(\neg\text{p})\}$
b. Didn't Jane go too?
$\{\text{FOR-SURE-CG}_x(\text{p}), \neg\text{FOR-SURE-CG}_x(\text{p})\}$

One major problem that arises for the VERUM approach to NPQs occurs with the interpretation of negative answers (Romero, 2006, Reese, 2007, Gutzmann and Castroviejo Miró, 2009). According to the denotation in (42b), a *no* answer to (41b) should pick out $\neg\text{FOR-SURE-CG}_x(p)$. That is, the speaker is not sure that p should be put in the common ground. However, *no* answers generally do not have this level of uncertainty attached to them. Instead, they seem to simply pick out $\neg p$ in such dialogues.

(43) a. B: Didn't Jane go too?
b. A: No.
$\rightarrow$ Jane didn't go
$\not\rightarrow$ B is unsure whether p should be in the common ground.

Romero (2006) suggests a possible solution to this problem would be to analyze VERUM as an expressive item. This would allow the contribution of VERUM/*really* to escape embedding like other expressives, e.g. *obviously*. Sure enough, in (45b) the fact that Kate didn't show up was not caused by the speaker's certainty. Similarly, in (46), the B's disagreement targets the embedded proposition ('That professor is very smart'), not the speaker's certainty on the matter.

(44) John is upset because obviously Mary doesn't love him.

(45) a. Kate didn't show up because she really couldn't make it.
b. John turned Mary in because Mary IS an alien.

(46) a. A: That professor really is very smart. / That professor IS very smart

b. B: That's not true.

Romero argues that answers to NPQs also pattern like this: *no* answers target the embedded proposition, side-stepping the VERUM component. However, it is not clear where this leaves the scope ambiguity that motivated the VERUM/NPQ analysis in the first place. Expressives are generally understood to act on a different semantic level to normal semantic content (Potts, 2007). So, it is unclear how VERUM could scope under negation, as is necessary for the ONPQ case. Problems also arise if we look at VERUM with respect to various presupposition holes, which expressives generally escape through. Here, we see that the speaker certainty associated with VERUM by R&H is diminished in conditionals and, indeed, in questions. So, VERUM does not act like other expressives if we take expressives to be conventional implicatures.

(47) p = John DID steal the files.

a. John DIDn't steal the files. (negation)
$\rightsquigarrow$ speaker is certain that $\neg p$

b. If John DID steal the files, we have to go to court. (conditional)
$\not\rightsquigarrow$ speaker is certain that p

c. DID John steal the files? (question)
$\not\rightsquigarrow$ speaker is certain that p.

However, this does not mean that VERUM is not an expressive item. It may simply be that the speaker certainty definition attributed to it by R&H is incorrect. This approach is taken by Gutzmann and Castroviejo Miró (2009) who argue that VERUM acts on the level of conventional implicature, where it interpreted as a separate performative. That is, VERUM(p) is an instruction to downdate the QUD, i.e. to add p to the common ground. Crucially, this presupposes that ?p is maximal in the QUD. The use of VERUM, therefore, produces two propositions. For the assertions, this produces the 'normal' assertion itself and the instruction to remove the asserted proposition off the QUD (a *use*-conditional meaning), e.g. (49). This indirectly provokes the feeling of speaker certainty associated with the use of VERUM. Similarly for NPQs, pre-posed negation asks the hearer to downdate p (a opposed to $\neg p$ (50). This makes the interpretation similar to the FOR-SURE$_x$ partitions on R&H's analysis in terms of producing bias.

(48) (Gutzmann and Castroviejo Miró, 2009)
$[\![\text{VERUM}(p)]\!]^c$ = The speaker c_s wants to downdate ?p from the QUD.

(49) John DID steal the files.
$\rightsquigarrow$ assert(p), VERUM(p), where p = John stole the files.

(50) Didn't John steal the files?
$\rightsquigarrow$ question(p), VERUM(p), where p = John stole the files

The CI approach solves the answer problem for NPQs in the way suggested by Romero (2006), i.e. by placing the extra meaning component in another meaning dimension. To account for the INPQ/ONPQ distinction in this way, we need to assume that p may be negative or positive. That is, the polarity of the clause embedded under VERUM determines the bias, since that is what the speaker asks the hearer to downdate. This means that the ambiguity of meaning here is no longer about scope.

However, there are several issues that remain if we adopt this approach to VERUM. First it is again not clear whether this definition of VERUM really patterns like other CI/expressives.[4] For example, VERUM$_g$ does not appear to be speaker-oriented in the same way as other expressive items. In (51) the prototypical expressive, *bastard*, cannot be attributed to the subject of the main clause, but instead is attributed to the speaker. In (52), however, the continuation clearly indicates that the speaker does not wish to downdate 'John stole the money' into the common ground, without being contradictory. In that case, the polarity focus seems to attach to Julia's beliefs about John stealing the money rather than the speaker's. Similarly, in (53) the speaker can question the antecedent of the conditional without contradicting herself or sounding like she has had a change of opinion.

(51) a. Sue believes that that bastard Kresge should be fired.
b. #But I think he's a good guy.

(52) a. Julia believes that Mary left because John DID steal the money.
b. but that's not evidence and I still don't think he did it.
$\not\rightsquigarrow$ speaker wishes to downdate John stole the money.

(53) A: Julia thinks that Mary left because John threatened her.
B: Julia has it in for John. Mary could have left for many reasons.
B: I don't think he would have done that.
A: Sure, but if Mary left because John DID threaten her, then we have whole other situation to deal with.
A: So the question is, did John threaten her?

Another issue with the mechanics of this approach is that p doesn't really need be

[4]One could make the argument that use-conditional meaning is just a separate thing from CI/expressives of the Potts type and hence things in this plane do not have to have the same properties as CIs. However, it seems clear that G&C do intend equate use-conditional and expressive meaning: '[W]e will treat VERUM in general as an expressive or, as we prefer to call it, a use-conditional item.' Gutzmann and Castroviejo Miró (2009, pg. 13)

maximal in the QUD for verum focus to be licensed. For example, in (54), Alice's utterance seems licit given that whether or not Mary is an alien had been considered (or even resolved negatively) by Alice and/or Bert, and that new information had come to light which resolves that question.

(54) Alice and Bert see Mary trip and cut her knee. Green blood gushes out.
A: She IS an alien!

In this case, we might say that the question, 'Is Mary an alien', is accommodated onto the top of the QUD (MAX-QUD) for Alice and Bert. However, in cases like (55), verum focus is fine even if *John stole the money* is already resolved/uncontroversial with no actual polarity change necessary.

(55) A: Is John an honorable guy?
B: He DID steal the *money*... But he did it to save Mary.
⇝ assert(John stole the money), downdate(John stole the money)
⇝ presupposes that John stole the money is MAX-QUD.
A: Right, right.

Here we need to assume that 'whether John stole the money' is added to the QUD so that the VERUM presupposition is satisfied. However, it is not clear how this differs from the normal act of assertion since, if we take the Ginzburg/Farkas view (cf. Section 2.3.3), assertions put a proposition on the QUD with the understanding that the speaker wishes to downdate it (i.e. have it accepted). Instead of being MAX-QUD, the constraint seems more to be that the positive truth value of the VF proposition be relevant to the task at hand. For example, determining if John is honorable in (55).

The main insight of the CI approach is that verum focus acts at the level of discourse structures like the QUD rather than the at-issue dimension, as do Potts's (2007) expressives which 'function solely at the contextual level, by actively changing the context'. However, given its lack of speaker orientation, its not clear that verum focus really belongs with the expressives as opposed to some other non-truth conditional dimension of meaning, e.g. the dimension where information structure acts. Kratzer (2004) suggests that infelicity arising from inappropriate focus marking patterns like inappropriate use of expressives. Again we are led back to the examples like (56) which gives a typical case of focus induced incongruity: the IS structure of the answer assumes that *someone ate the beans* is given, which is not the case for this dialogue. The verum focus case patterns in the same way in that it suggests that 'Fred ate the beans' is given, i.e. in the conversational background.

(56) A: What did Fred eat?
B: FRED ate the beans.
B: Fred DID eat the beans.

So, at this point it seems there is good reason to think that verum focus provides information outside of the at-issue or proffered content of the utterance. However this information seems to act on the same dimension as narrow IS focus in general and not that of expressives.

5.3.3 Arguments Against a Special Treatment of Verum

So why wouldn't we want to fold our analysis of VF into the more general treatment of IS focus? Gutzmann and Castroviejo Miró (2009), following Höhle (1992) claim that VF cannot be reduced to informational or contrastive focus, and hence requires a special semantic operator. Verum's specialness comes from the fact that the truth value is focused independently of the meaning of actual lexical item that is made prominent. This is apparent in the German data, where VF can be triggered by prominence on a finite verb, an auxiliary, or a complementizer. What is important is that the item that is stressed is syntactically in the C domain.

(57) A: Hanna claims that Carl is writing a screenplay. (G&C:7)
B: Karl SCHREIBT ein Drehbuch
Karl writes.VF a screenplay
"Carl IS writing a screenplay."

(58) A: Hanna claims that Carl has written a screenplay. (G&C:8)
B: Karl HAT ein Drehbuch geschrieben
Karl has.VF a screenplay written
"Carl HAS written a screenplay."

(59) A: Will Peter stop smoking? (G&C:13)
B: Peter HÖRT mit dem Rauchen auf
Peter stops.VF with the smoking PART
"Peter WILL stop smoking"

(60) A: Peter talks as if he were a philosopher. (G&C:17)
B: Ich denke DASS er ein Philosoph ist
I think that he a philosopher is
"I think that he IS a philosopher."a

Focus as Emphasis

The data here is congruent with the view that this operator simply expresses focus on some feature (that ends up) in the C domain that cannot bear stress on its own, e.g. positive polarity. That is, we assume that German does not have a 'focus' pronunciation for positive polarity like English *do* (c.f. Hyman and Watters, 1984, for Bantu) and is so parasitic on whatever ends up in the appropriate syntactic position. So, these facts do not actually seem to be too much of a barrier to bringing VF into the focus fold. Gutzmann and Castroviejo Miró's (2009) argument against this is basically centered around the fact that VF behaves differently to 'normal' content word focus in some situations. With respect to the data above, G&C seem to assume that one job of focus is to emphasize the content of the focus carrier. This content emphasis doesn't arise with VF where, for example, 'it is far from clear what content could be emphasized by stressing the complementizer' (pg. 6). However, although focus prominence can provide emphasis, from the informational viewpoint of this prominence is to indicate kontrast and mark IS status. So, the fact that the phonological carrier of prominence is in the right position is enough to signal polarity alternatives (in the right context), which fits with the pattern of the data.

Focus changing truth values

The second difference that G&C note is that the position of focus with respect to content words can change truth values in the presence of a focus sensitive operator like *only*, e.g. (61). However, the difference between utterances which minimally differ with respect to VF is not about truth values.

(61) John introduced Bill to Mary, James to Sally, James to Mary.
  a. John only introduced Bill to *Mary* $\rightsquigarrow$ True
  b. John only introduced *Bill* to Mary $\rightsquigarrow$ False

(62) a. Mary ate a cake.
  b. Mary DID eat a cake.

However, just as we do not want to identify focus with emphasis, we do not want to assume that information focus is in and of itself about truth conditions. Given that we take VF to be focus on (positive) polarity, it's not very surprising that its presence does not effect truth conditions just as it is not surprising that focusing *Mary* in (61) does not change the fact that *Mary* appears in the proffered content[5] of the

[5]i.e. in the ordinary semantic value of the utterance.

utterance and not someone else. Moreover, without a focus sensitive operator, focus does not change the truth value of the proposition. What it does change is whether the utterance is a felicitous thing to say in a given situation, as in the following (63).

(63) What did Mary eat?
   a. Mary ate a CAKE
   b. #MARY ate a cake

Here we see question answer congruence coming into play. In fact, this is much the same as the MAX-QUD presupposition G&C associate with verum focus. The difference is that VF is congruent to polar questions, while the DP focus answers a wh-question. So, we can see that VF doesn't act like content word informational focus precisely because the alternatives that VF generates are about polarity. To illustrate, if we were to try to apply *only* (in the manner of Rooth (1985)) to the VF utterance, we get the interpretation that p is true and no other truth value holds of it, i.e. (64c). This must always evaluate to true so trying to apply *only* here is just vacuous.[6]

(64) Mary only DID eat the cake.
   a. $[\![\text{Mary only did eat the cake}]\!]^o$ = Mary ate the cake
   b. $[\![\text{Mary only did eat the cake}]\!]^f$ = {p=Mary did eat the cake, Mary didn't eat the cake}
   c. $\forall q[q \in \{p, \neg p\} \text{ and } \hat{} q] \rightarrow q = p]$

Overall, the arguments that G&C bring up against reducing VF to informational focus derive from an assumption that focus is characterized by the behaviour of emphasized content words rather than the more general information packaging notion of focus. In general, the differences between VF and other forms of informational focus are about the domain of that VF marks, rather than what it does in terms of information structure.

[6] *Only* above VF is good if it associates saliently with something lower. For example, in the following it associates with ten miles an hour.

(i) Even if I only did go ten miles an hour, I had a blast. `http://princesserin.tumblr.com/`

(ii) a. You shouldn't be so hard on Mary. At least she didn't raid your entire food store.
   b. Yeah. I guess, she only did eat the cake.

5.3.4 The IS View of Verum

Under an IS view, VF simply marks the IS focus of an utterances as being the (positive) polarity of the at-issue proposition. Thus, the topic is the core proposition itself and so proposition is already in some sense discourse old. Like other IS focus types, VF is constrained by QAC. To achieve congruence, VF requires a a polar question. One situation where this configuration is desirable is when a speaker wants to update the polarity of a proposition already in the background. This happens, as previously mentioned, in direct answers to polar questions, new information updates, as well as direct affirmations and contradictions, e.g. (65).

(65) A: I don't think Marianne met with Lenny.
B: No! Marianne DID meet with Lenny.

Looking at the data this way, we see that VF performs the role G&C seem to give informational focus (i.e. the answer to a question). However, IS focus is inherently contrastive – the actual content contrast with alternative updates that could have been made. In fact, it is not too hard to find examples where VF is used contrastively. This is something not predicted by the CI analysis of VERUM. VF naturally appears when we want to contrast polarity across (sets of) propositions. For example, in (66), we can think of each conjunct as answering a polar question. Here, the positive or negative polarity of the answers contribute to the resolution to a bigger question (i.e. whether the speaker should get paid or not).

(66) I didn't do the dishes, but I did do the laundry and I did mow the lawn. So you should really pay me!

Examples where VF is used in indirect responses can also be framed as cases of polarity contrast between sets of propositions. In (67) B's response can be taken as marking a piece of evidence that supporting A's claim.

(67) a. A: I think Marianne has a crush on Lenny
b. B: She DID buy him *Valentine's day card...*

That is, if we consider a check list of possible indicators that Marianne has a crush on Lenny, Marianne buying Lenny a card gets check mark. So, the speaker in both (66) and (67) seems to be employing the notion of discourse strategy: answering subquestions in order to answer a question higher up in the discourse tree (Roberts, 1996, Büring, 2003). What makes VF responses different from other responses is

that they project polar questions (*whether* questions) in terms of question/answer congruence. A non-VF declarative may invoke other sorts of questions.

(68) a. A: I think Marianne has a crush on Lenny
b. B: She bought him a *Valentine's day card*...
⇝ What did Mary buy Lenny?
⇝ What did Mary do?

Now, in the indirect VF response cases like (67) we clearly get a clash between MAX-QUD and the polar question projected by the VF utterance. Our hypothesis is then that this breach of question/answer congruence is licit only if the VF speaker signals that they are playing a strategy. That is, in order to be co-operative they will attempt to resolve the dominant question (usually MAX-QUD) by answering subquestions. This is reflected in the second prominence which seems necessary in indirect VF responses, which marks another unit as contrastive. Since the inserted *do* is the IS focus, the second prominence signals a contrastive topic, in the sense of being an alternative generator in the ground. For VF, this seems to project an evidence checklist. For example, for (67), a possible strategy iterates over the VP: *whether Mary bought Lenny a Valentine*, *whether Mary flirted with Lenny at the party*, etc.

The idea that VF projects implicit polar questions in the discourse tree is very similar to G&C's idea that VF(p) requests a downdate of $?p$ from the QUD. However, we do not need to posit any extra machinery than we already need to deal with IS focus and kontrast/F-marking.[7] In the end, the difference between VF and focus on content words is parallel to the difference between polar questions and wh-questions.[8] The difference in both cases stems from the sorts of alternatives the different forms generate. An information packaging approach also says more about why you would want to use VF in the first place. You want to use it when you want to say something about polarity. Sometimes you want to say something about polarity because the status of a proposition has changed in that respect. Sometimes you want to say something about polarity because considering the polarity of a proposition is a relevant piece of a strategy attempting to answer an unresolved question.

[7] It's entirely plausible that Gutzmann and Castroviejo Miró's (2009) notion of use-conditional semantics could be used to do some heavy lifting with respect to IS focus in general, but this is beyond the scope of this project.

[8] Clearly, we would prefer to have a theory of questions that encompasses both polar and wh-questions, so why make the difference at the declarative end?

5.4 Production Experiment: Declarative Response Prosody and Verum Focus

5.4.1 Goals and Expectations

Now that we've established what verum does in informational structural terms, we can start moving other parts of the discourse structure about to get a better idea of the relationship between the interpretation of these utterances and the presence of terminal rises. In the discussion above, we identified two main response types that have the same basic dialogue move type, statement/assertion, but are intuitively pronounced with different intonational forms. On the one hand, direct responses seem to require a strong prominence on the inserted *do* followed by post-focus compression, i.e. a falling nuclear accent. On the other hand, indirect responses seem to require a post-focus prominence and a low rise, i.e. a fall-rise accent. So, it seems that these sorts of response dimensions may be more consistent predictors of meaningful prosodic variation than the usual type of dialogue act. However, to support this idea, we need to test how robust these intuitions are. Furthermore, we would like to know whether the intonational form is affected by whether a response is an agreement or contradiction.

In order to control for these factors, a production experiment was designed to elicit verum focus declaratives in various contexts. The primary goal was to test whether the different response types within a dialogue act category (direct/indirect, agreement/contradiction statements) mapped to well defined intonational categories. In particular, we want to see if indirect response contexts elicit terminal rises. The amount and types of variation in the data will have implications for current analyses of intonational meaning, not just with respect to terminal rises, but the interpretation of pitch accents, specifically the plethora of analyses of the fall-rise accent. Several such accounts claim that pitch accent shapes mark specific IS categories. For example, Steedman (2000) makes a distinction between ToBI L+H* and H*, wherein the latter marks IS focus, while the former bitonal accent marks an element in the ground. Büring (2003) doesn't make this distinction, but instead claims that falling accents (H*L%) mark focus while fall-rises ((L+)H*LH%) mark contrastive topics, with each IS category having different consequences for the logical form of the utterance. The basic data for his analysis come from pair-list answers to wh-questions.

(69) a. What about Fred? What did he eat?
b. *Fred*$_{\text{H*LH\%}}$ ate the BEANS$_{\text{H*L\%}}$

(70) a. What about the beans? Who ate them?

b. FRED$_{\text{H*L\%}}$ ate the *beans*$_{\text{H*LH\%}}$

In Büring's account contrastive topics are associated with partial QUD resolution (or at least an impression of such). This is formalized in terms of strategies: the speaker is answering a subquestion related to the actual QUD, and so doesn't resolve the QUD fully. For examples of the type above, the idea is that by using the fall-rise accent, the speaker signals that the superquestion of *who ate what* has not been completely answered by the response given, e.g. possibly Miranda ate cake.

This seems to fit the indirect response contexts we have seen above. If the distinction is correct, we would expect to consistently see a terminal rise in indirect response scenarios. Although previous experimental work indicates that these mappings are too tight, such studies indicate that IS does play a consistent role in determining the prosodic form of an utterance. For example, semi-spontaneous data from Calhoun (2012) indicates that the topic/focus difference should be characterized in terms of relative prominence instead of accent/boundary shape. However, while the design of that experiment did elicit topics (themes) that were in contrast with other elements of the discourse, it did not really address the same sort discourse configurations linked to Büring's notion of *contrastive topic*. Calhoun's study involved a two-player collaborative game where the participants (the Driver and the Slider) had to use prespecified frame sentences like the following.

(71) (Calhoun, 2012)

a. Driver: If we do this, will the banana land on some money?

b. Slider: No, the [lollipop]$_{\text{focus}}$ will land on some [money]$_{\text{topic}}$, the [banana]$_{\text{topic}}$ will land on a [monster]$_{\text{focus}}$.

Note that the Slider's first clause above provides a complete answer to the Driver's question, so there is no reason to mark the answer as part of a subquestion strategy. So, this sort of response seems to be better characterized as *direct*. The majority of topic tokens were not produced with a clear accent (71%) suggesting that they were deaccented. Impressionistically, it does seem that accenting the topic in those scenarios would be a marked option rather than the default. Intuitions about the verum data however suggest that second post IS-focus prominence is necessary for in the indirect responses. So, we would like to compare how reliably those discourse contexts elicit prominence in the IS ground.

Response Types	
D	Direct
I	Indirect
A	Agreement
C	Contradiction
Q	Question
Stimuli	
E	Emily did bring a meringue
M	Marianne did meet with Lenny
W	William ran away
Y	Mary remembered your birthday
B	You are in a bad mood
HA	Annemarie did wish you a happy birthday
S	She did just have surgery
P	You did say you liked it when we had it last year
WA	William did steal the money
T	Marianne did make some bad mistakes this time
MB	Marianne provided the building schema
LB	Lenny provided the password
BB	Bill got you a cashmere shawl
AB	Annemarie got you that tree fern

Table 5.2: Identifiers for production target sentences

5.4.2 Data and Method

The production experiment consisted of two parts, both recorded in pairs: two scripted dialogues (long ≈ 30 turns each) and statement/response pairs (short). The target responses in the latter task consisted of two declaratives with verum focus (E: 'Emily did bring a meringue', M: 'Marianne did meet with Lenny') and two with broad focus (W: 'William ran away', Y: 'Mary remembered your birthday'). Four distractor sentences were also included, involving *really* placement, e.g. 'Anna (really) is (really) in Rio'. The motivation for using these constructions was to examine the relationship between dialogue structure and prosody while keeping the IS fixed. As discussed previously, in verum focus sentences, IS focus is the proposition's polarity. This draws the main sentence stress from the default utterance end position to the inserted 'did'. So, unlike the broad focus cases we consider, post 'did' accents will mark units in the IS ground (or theme) rather than the IS focus. In broad focus configurations, the IS focus is the proposition itself. The pairs were designed to cover direct (D) and indirect (I) agreements (A) and contradiction (C) type responses, as well as declarative questions (DQ). This resulted in 20 ({E,M,W,Y}x[{D,I}x{A,C}+{DQ}]) target stimuli of similar form to (1)-(3), read twice by each participant (see Appendix A.1 for full list of stimuli).

The scripted dialogues involved two scenes where the participants talked about a past event. The scenarios were set up to elicit direct and indirect agreements and contradictions. Four turns reproduced conditions from the short context recordings (EIC, MDA, WIA, YDC). The distractor sentences and four other verum and broad focus target sentences were included in the dialogues. A pair list partial answer sequence was also included in each dialogue to give a more direct comparison to the type of data examined by Büring (2003) in his analysis of contrastive topic. (see Table 5.2, and Appendix A.2 for full dialogues.). Again, each speaker recorded each part twice. However on the second recording participants were asked specifically to try to sound more involved/engaged in the scene. All recordings were recorded in a sound attenuated booth. Eight pairs of speakers of Standard American English participated (7 males, 9 females) and were paid for their effort.

Timing data was initially obtained by using using the Penn Phonetics Lab forced aligner (Yuan and Liberman, 2008), after which word and turn boundaries were manually corrected. F0 and Intensity features were extracted via Praat. F0 contours were also manually corrected via Xu's ProsodyPro Praat script (Xu, 2011). As in the corpus studies in Chapter 3, these values were normalized to a semitone scale relative to each speaker's median F0. Intensity and duration measurements were converted to z-scores by speaker and word respectively. In the following analyses we look primarily at Legendre polynomial coefficients as describing height, tilt and convexity of F0 and intensity contours (cf. Section 3.3.2).

Hypotheses

- We expect nuclear accent placement in direct responses to be determined by the IS focus. That is, on the inserted *do* in the verum case and on the utterance final word in the broad focus cases. We don't expect to see prominences after the IS focus in the direct verum cases. We do expect to see a another prominence after the IS focus in the indirect cases. This will be determined by what element generates the appropriate contrast in the context.
- In terms of terminal rises, we expect the declarative questions to exhibit convex nuclear rises from the IS focus to the utterance end. We also expect to see low rises in the indirect responses.
- We do not expect to see pitch accent shape (e.g. in terms of peak alignment or a following rise) as a distinguishing feature between IS focus and topic, although this is what is predicted by some theoretical accounts (Steedman, 2000, Büring, 2003).
- Given our previous findings on the prosody of *really*, we do not expect contradictions and agreements to be associated with specific pitch accent targets (contra Steedman, 2007). We do expect pitch range and pitch height to increase with speaker involvement.

5.4.3 Short Context Productions

Figure 5.4 shows mean F0 value (in semitones) with respect to normalized time for each of the target sentences. From inspection, it appears that most of the variation between conditions happens at the end of the utterance. Direct responses generally fall to a much lower value than indirect responses. As expected, the declarative questions are generally rising through the final word, though there is relatively less of a rise in the verum sentences because the F0 level is already high from the rising accent through 'did'. The indirect response show a fall-rise shaped accented on the final word for sentences E, W, and Y. For sentence M, the fall-rise is stretched from the word 'meet' to the end of the utterance. This is as expected since the contrast falls on the verb for the M the context and the final word 'Lenny' is given in the previous utterance, while it falls on the final word for the other three sentences.

(72) A: I think Marianne might be conspiring with Lenny
B: Marianne DID *meet* with Lenny

There does not appear to be much of a shape difference based on whether response was an agreement or contradiction at the tail end. The exception to this is the Y direct

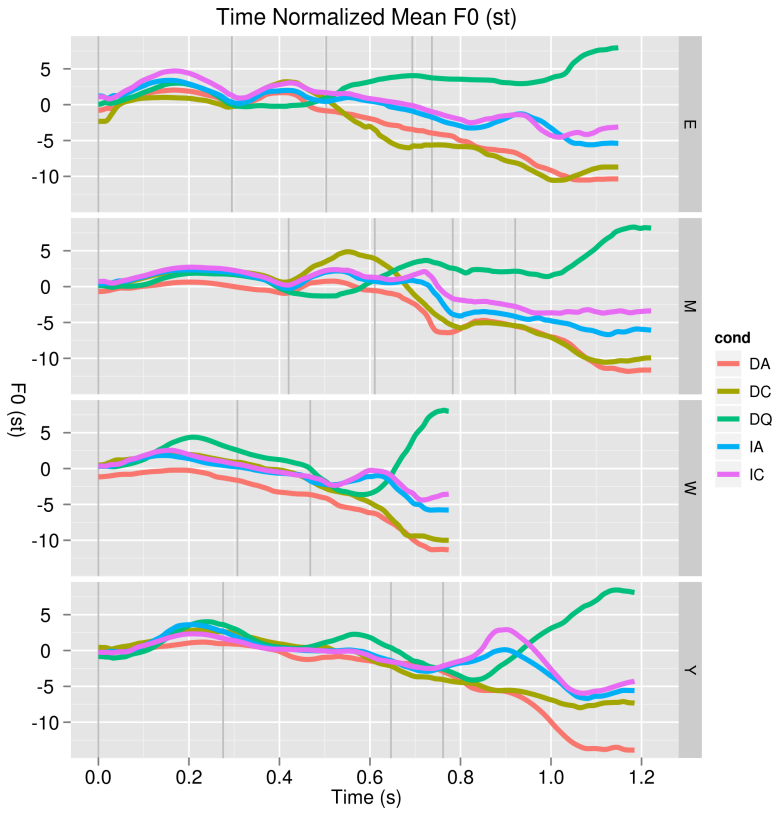


Figure 5.4: Time Normalized Mean F0 (st) for short context productions. Vertical grey bars indicate word boundaries.

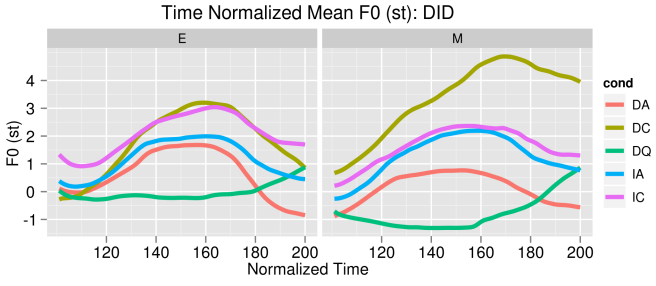


Figure 5.5: Verum focus 'did': Mean F0 (st).

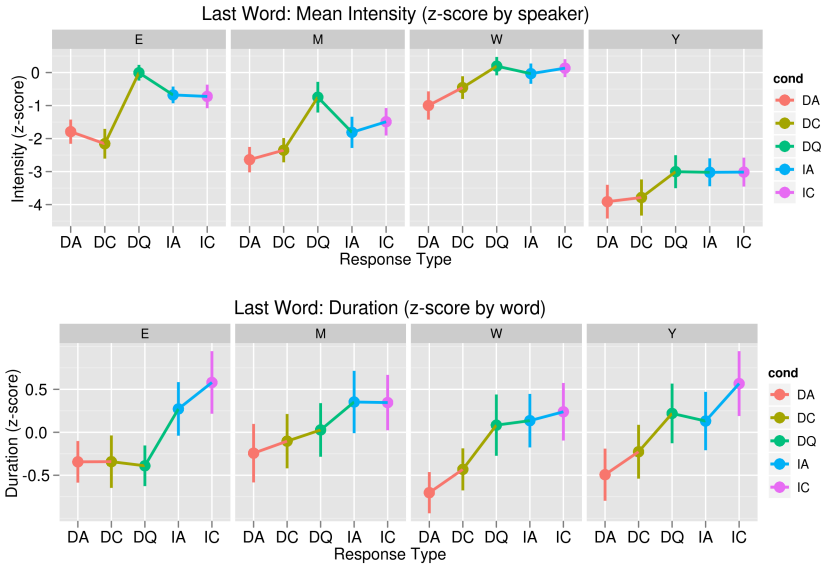


Figure 5.6: Last word intensity and duration by response type

contradiction ('Mary remembered you birthday') which seems to not fall as much as its agreement counterpart. If we look at the inserted *did* (Figure 5.5), however, we can see that the direct contradiction is higher than that of the direct agreement, and that this is more the case for the M than the E sentences. We see much less variation by condition for the mean intensity for the last word. However, it does appear that the indirect and declarative question productions have a higher intensity. This is not surprising assuming an extra prominence in the indirect responses. In the same vein, we see relatively longer final word duration for the indirect and question responses (Figure 5.6).

F0 Shape Features

Figure 5.7 shows the distribution of the data in terms of F0 height, tilt and convexity. In order to capture the fact that the last prominence does not necessarily fall on the utterance final word, we show the fitted values for the first three Legendre coefficients from the position of the last contrast invoked in the indirect case to the end of the utterance. For this data set, this basically means taking the measurement for sentence M ('Marianne did meet with Lenny') from the word 'meet', and from the last word for the other sentences. To aid interpretation of these projections, Figure 5.8 shows the F0 contours generated from mean Legendre coefficient values. As we expect, declarative questions have positive values for the second Legendre coefficient reflecting their rising shape. The indirect responses have greater concave fall, evinced by the

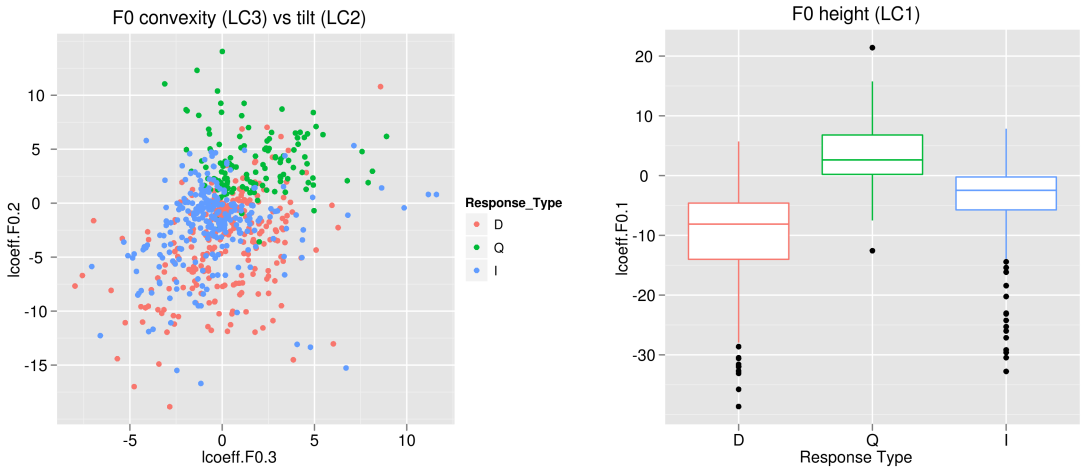


Figure 5.7: Shape Features: Comparing position of last contrast

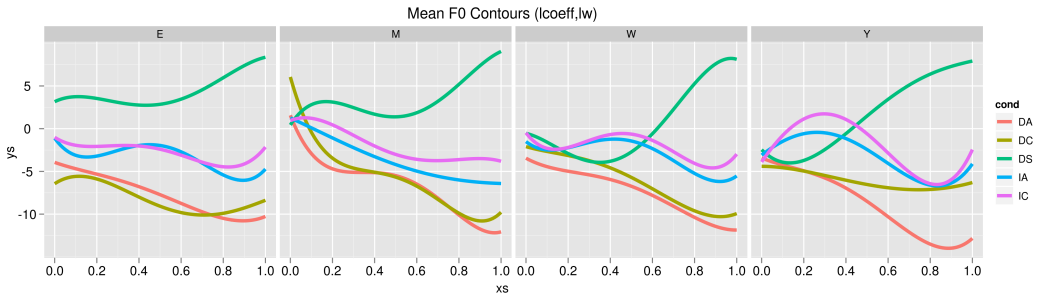


Figure 5.8: Mean F0 Contours based on Legendre coefficients, last contrast

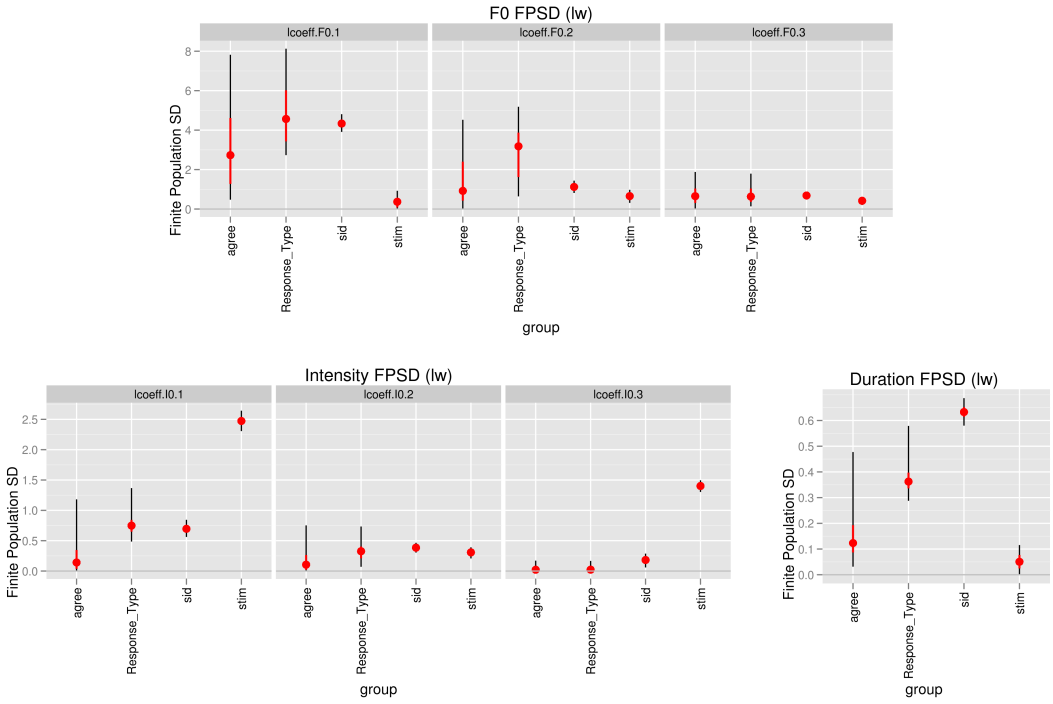


Figure 5.9: Analysis of variance: Finite Population Standard Deviation (lw)

negative values for lcoeff.f0.3. In general, the F0 height is higher for the indirect responses than for the direct ones. However, we see quite a lot of overlap on all of these features for the statement type responses.

Category Differences

Multilevel linear regression was used to examined the effects of response type (direct, indirect, question), agreement (agreement, contradiction, question), stimulus, and subject on height, tilt and convexity of the F0 and intensity contours as well as z-score duration. In the following, we model these as group level predictors. Figure 5.9 shows the finite population standard deviations of the different factors for the utterance end (last contrast). This indicates the amount of variation in the data explained by each factor. Unsurprisingly, we see significant amount of speaker variation for all the features. Factoring this out, it appears that response type accounts for a non-zero amount of variation for the different prosodic features except intensity convexity, while the agreement type only really seems at play for F0 height.

Figure 5.10 shows the parameter estimates for response and agreement types for the F0 shape features. We see a clear significant difference in the effect of the response types with respect to F0 height: direct responses are lower, questions are higher, while

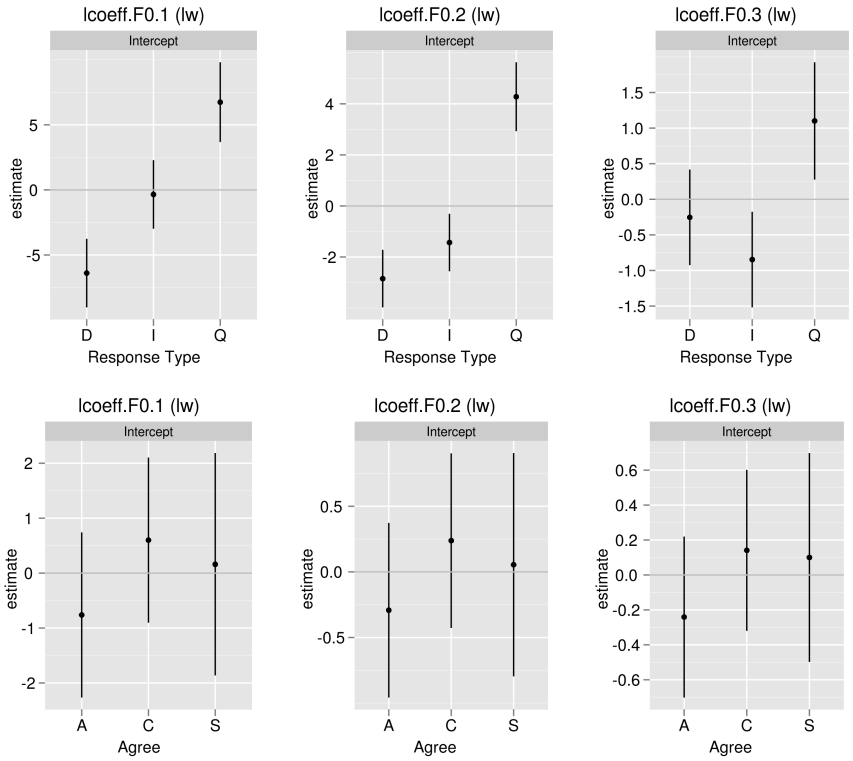


Figure 5.10: F0: Response Type

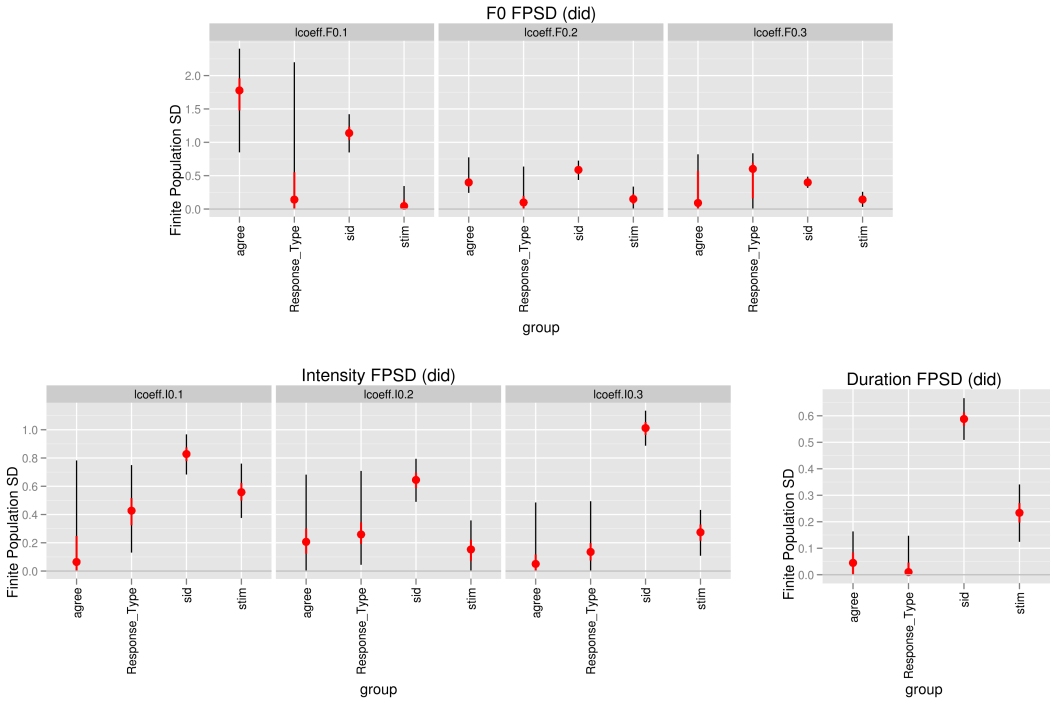


Figure 5.11: Analysis of variance: Finite Population Standard Deviation (verum/did)

being an indirect responses puts you in between. In terms of tilt and convexity, we see the trend is for indirect responses to have a less falling tilt and a more concave shape (i.e. to be more peaked), but the difference between these two categories is not as clear as for F0 height. Intensity and duration features pattern in a similar way with indirect statement responses exhibiting higher mean intensity and duration than direct responses, and questions have even higher values for those features.

This supports the idea that indirect responses exhibit a second prominence. While it seems that this often manifests in a rise-fall-rise shape, not all are actually terminally rising (after the accent peak). A negative slope through the last 100ms was found for 46% of indirect responses. So, the rise component is better characterized as being less of a fall. This casts doubt on the idea that contrastive topics are characterized by a high boundary tone. We don't see any clear differences between the agreement/contradiction/question categories, although on average direct contradictions to fall less steeply than direct agreements (mean DA: -28.40 st/s, DC -18.02 st/s).

Figure 5.11 shows that the agreement dimension accounts for more variation in the data when we look at the verum triggered *did* in sentences E and M. The parameter estimates in Figure 5.12 indicates that contradiction versions of *did* were higher in F0 than the agreements, confirming the impression that DC productions were 'bigger'

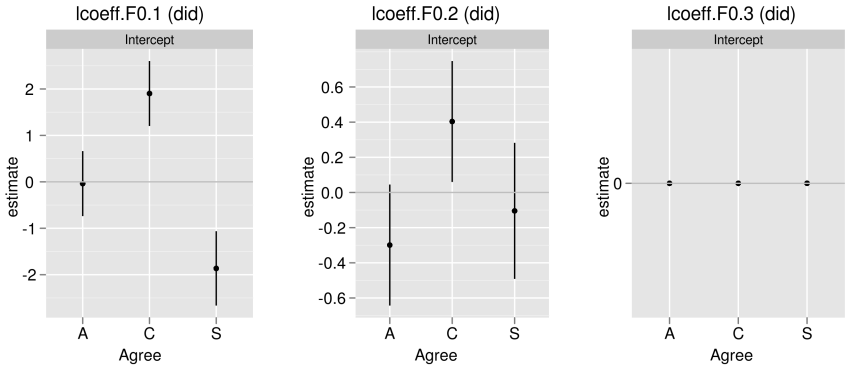


Figure 5.12: F0: agreement (did)

with respect to F0. In fact, we see that both F0 tilt and relative time of the maximum are higher for the DC class than the DA class, which is indicative of peak delay. This goes against the hypothesis that disputed IS units should be marked with L* type tones, as the DC shape is clearly neither low nor falling to a low target. Instead, the contradiction accent employs gestures that are more emphatic in the manner of Gussenhoven (2002) who find that delayed peaks are perceived as having a larger pitch range and are hence more effortful. Similar to the F0 features, we do not find significant differences in the intensity or duration features with respect to the agreement/contradiction dimension, although the trend is for contradictions to be longer.

Speaker Differences

The parameter estimates for the prosodic feature models indicated significant speaker variation. We can get an idea of the ways speakers vary by directly comparing their productions. Figures 5.13 and 5.14 show speaker differences for utterance final productions of the word 'meringue'. We see the expected separation of indirect and question responses from the direct responses in terms of, F0 and intensity height, as well as an FR pitch accent shape in the majority cases. However, we also see quite a stark variation in F0 range, with the contours of speakers jj1, ma2, sn1, sn2 all appearing quite compressed. Notably, these compressed range speakers were all male.

To factor out this pitch range compression, we examined F0 contours for these speakers transformed to z-scores based on the utterance mean and standard deviation. The results in Figure 5.15 now show more of a FR shape for jj1 and sn1. While the z-score data don't show much more of an FR accent shape for speaker sn2, we do see more of spread of corresponding intensity values in Figure 5.14 (similarly for sn1,

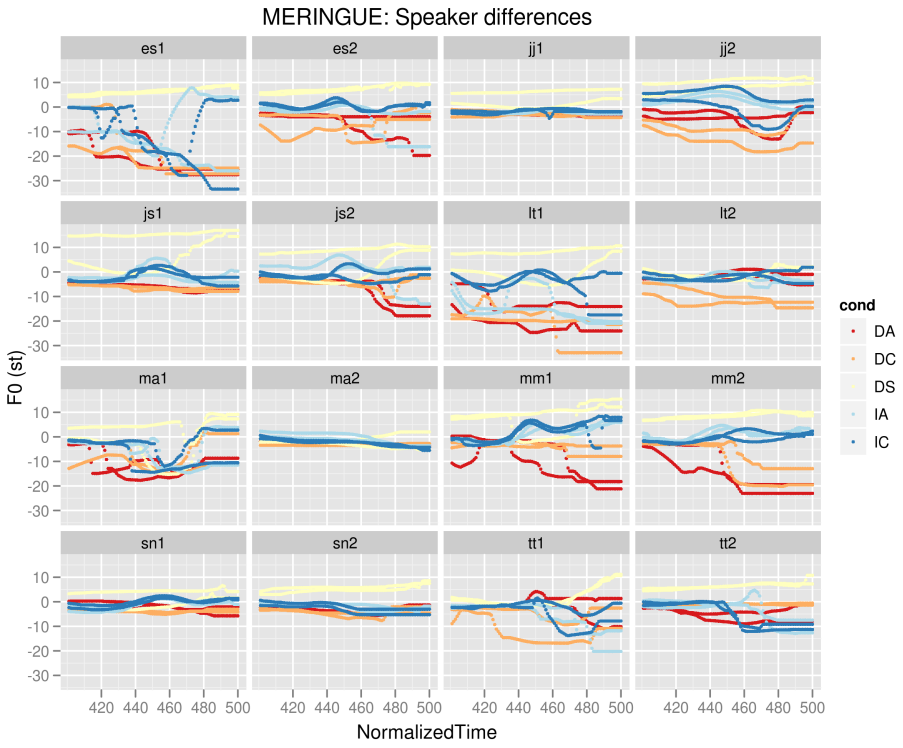


Figure 5.13: F0: speaker differences for utterance final 'meringue' (sentence E)

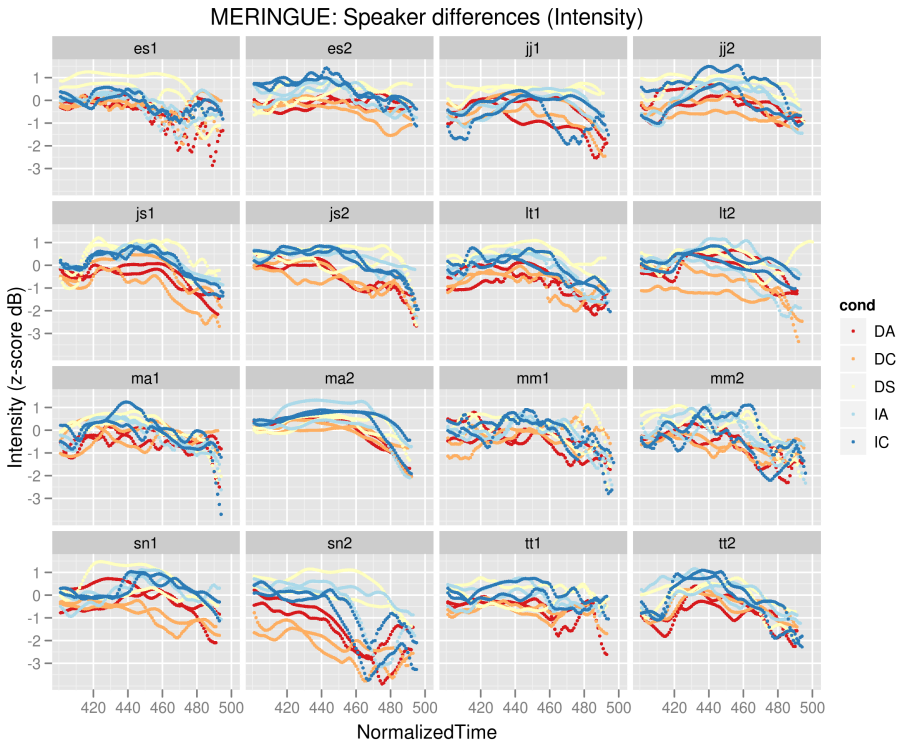


Figure 5.14: Intensity: speaker differences for utterance final 'meringue' (sentence E)

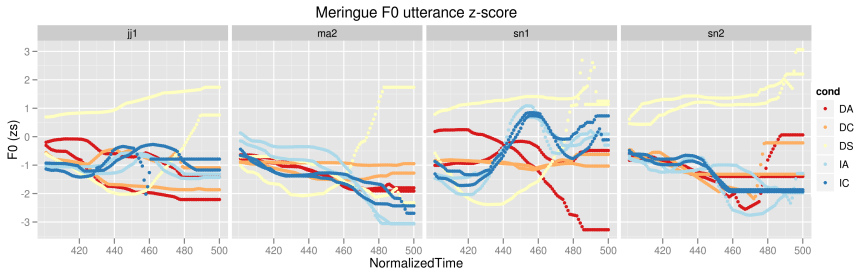


Figure 5.15: F0 z-scores by utterance: 'meringue' (sentence E)

and to a less extent for ma2). Given the relative compression of the corresponding F0 contours, it seems that some speakers are using intensity more than F0 to signal a second prominence on the indirect responses. This supports the idea that it is overall prominence that is important for this sort of response type rather than the specific pitch accent shape.

Note, that the discontinuities in the F0 contours reflect creaky phonation. While almost all the speakers showed some level of utterance final creakiness for the direct responses, this non-modal phonation was especially pervasive in the productions of speaker es1. However, onset of creaky voice was later for the indirect responses than the direct responses for this speaker where it seemed to start immediately after the verum 'do'. As with the intensity differences discussed above, this suggests that differences in voice quality were being employed as a signal of deaccenting (or conversely modal voice signalled non-deaccenting).

5.4.4 Long Context Productions

Figure 5.16 shows the mean time normalized F0 contours for the long context data for the stimuli that appeared in the short productions. The means for the long context productions match the contour expected with respect to the short context data given the response type. In particular, we see a fall-rise shape on the indirect responses (E, W). The direct contradiction (YDC), however, ends relatively high. This seems to be a result of the use of the contradiction contour (Liberman and Sag, 1974) in this context. This contour is characterized by high utterance initial pitch followed by a low section and a terminal rise.

The use of this contour varied between and within speakers. Figure 5.17 shows the YDC contours for two speaker's ma2 and mm1. Speaker mm1 consistently used the contradiction contour in both short and long productions, while ma2 used it only once on the second repetition of the dialogue, i.e. the more 'intense' version. This

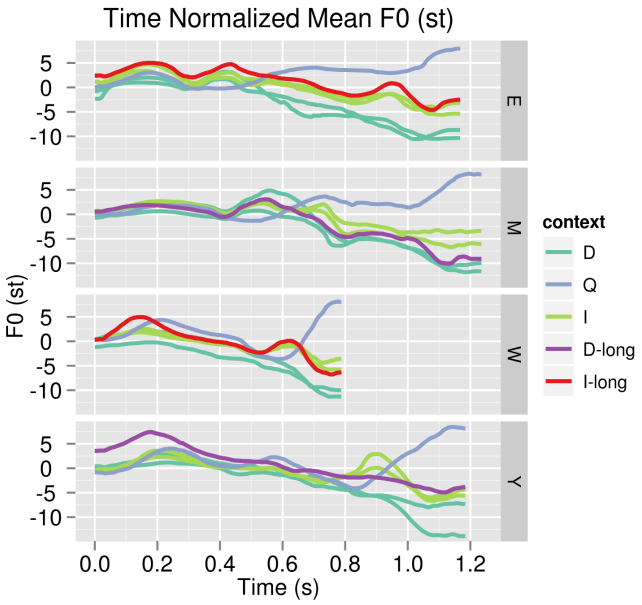


Figure 5.16: Long vs Short Contexts: Mean F0

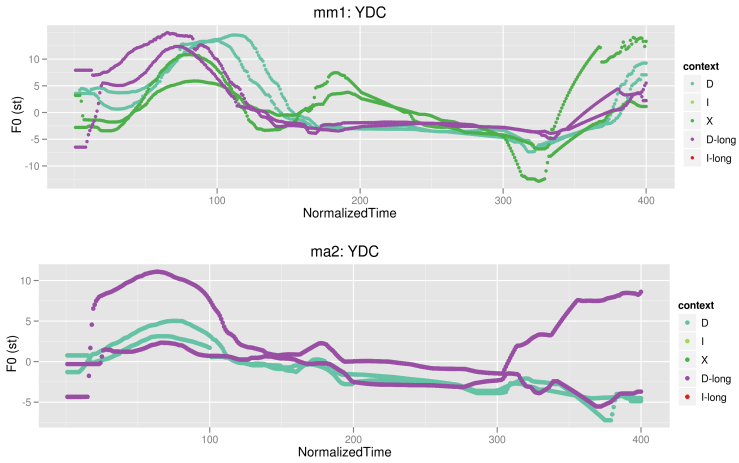


Figure 5.17: Contradiction Contour

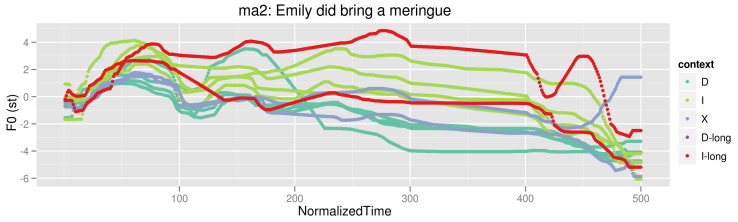


Figure 5.18: Long and Short: ma2, sentence E

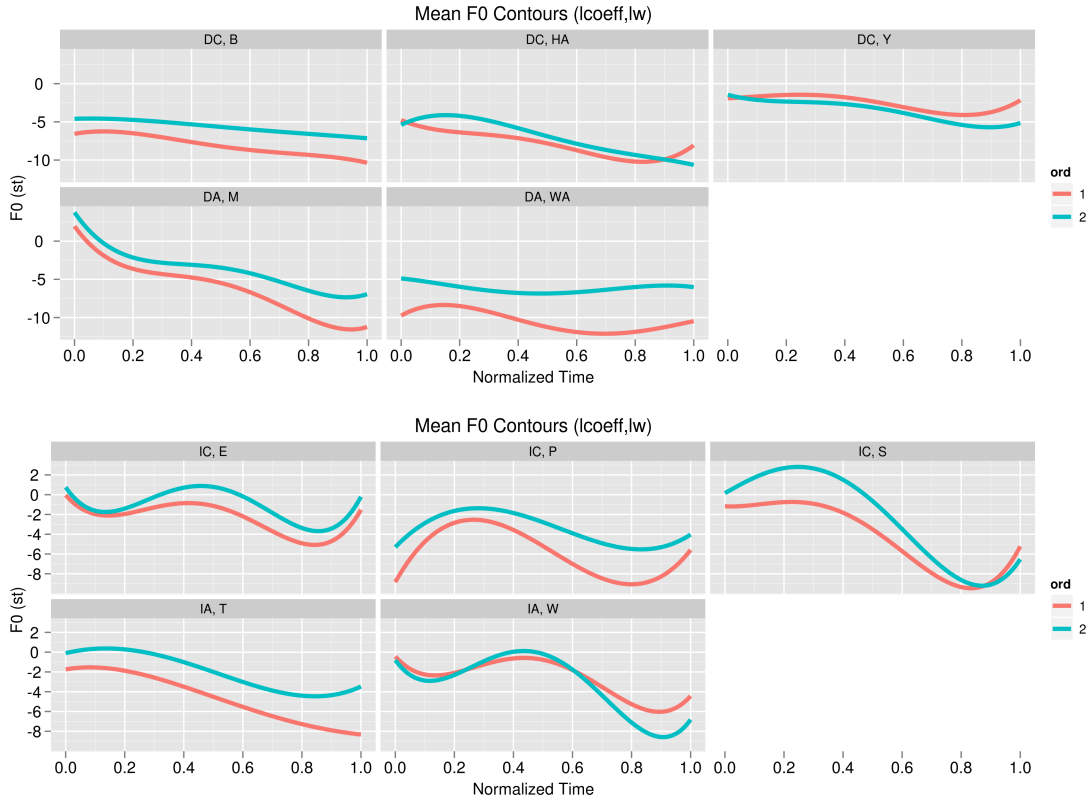


Figure 5.19: Direct (D) and indirect (I) verum responses in the long contexts

speaker similarly produced more of an accent in the second production of sentence E (EIC, Figure 5.18).

Overall, the indirect responses exhibited higher F0 level and a more peaked shape than the direct response productions. This provides further support for the idea that the former configuration requires a second prominence. Generally we see that the more involved the speaker is (or at least intends to sound) the bigger and more effortful the intonational gestures appear. As the gestures balloon, we get a corresponding increase in pitch range and pitch level. Figure 5.19 shows the mean boundary F0 contours by repetition order for the additional verum responses included in the longer dialogues. The graphs show that the second more involved reading again had higher F0 values on average. The intensity data shows the same trend. However, while we do see a trend, we do not find the effect of reading order to be a significantly different from zero, adding this factor to the multilevel regression of the form used in the analysis of the short context data. This is to be expected given that there is some inherent variation in how involved the different speakers were in the task. So, perceptual data seems necessary to give a more sound idea of how prosody varies

with this sort of affect. Nevertheless, it seems to be a fairly safe bet that increasing overall F0 and intensity levels will be interpreted as a signal of greater speaker involvement/engagement.

5.4.5 Functional Principal Components Analysis

Thus far we have used Legendre polynomial decomposition to characterize the height, tilt and convexity of our prosodic contours. It seems clear that each of these shape dimensions has some value in characterizing the differences between the discourse categories we have been looking at. We would like to quantify further what sort of variation exists in the data with respect to these sorts of shape features, and similarly where the locus of this variation is. As such, we employ Functional Principal Components Analysis (FPCA) to get a better a summary of how the shape of these utterances vary as a whole (Ramsay, 2006). Like regular PCA, we use this technique to describe our contours as a linear combinations of principal components. In the functional case, these principal components, or *harmonics*, are functions which represent the dominant modes of variation in the data. Like the Legendre polynomials, these functions can then be used as a basis upon which to describe the contour shapes of in our data set (cf. Section 3.3.2). Unlike the Legendre coefficients we have been looking at, this method identifies a set of basis functions which are tailored to the actual data set. We can use this to better pinpoint where the variation is happening in the contours.

We follow Gubian et al. (2010) in deriving the FPCA scores (i.e. harmonic coefficients). The data was initial fit using B-splines and FPCA was performed on the short context productions using the R package `fda`. As above, we looked at the segment from the last actually prominent word at or after the expected main stress position (fda.lw). So, for example, we take *'meet with Lenny'* for sentence M indirect responses, but only *'meringue'* for sentence E. This was done to capture the fact that the accent and tail may span multiple words. We also look at the F0 contour over the whole of the utterance. Since the number of words in each utterance varied, the time scale of each utterance was warped so that each contour had the same duration and in addition matched at two landmarks: the end of the subject (actually the topic) and at the last words point (e.g. 'meet' for indirect response M). The results of this are shown as feature set fda.warp.

Figure 5.20 shows the mean function (means of basis function coefficients) and harmonics for the whole utterance (time warped) data set. The shape of the harmonics in the latter set confirm that most of the variation is indeed happening towards the end of the utterance, i.e. after the second landmark. Figure 5.21 shows the mean and harmonics (principal components) for the last words data set. Here, the first

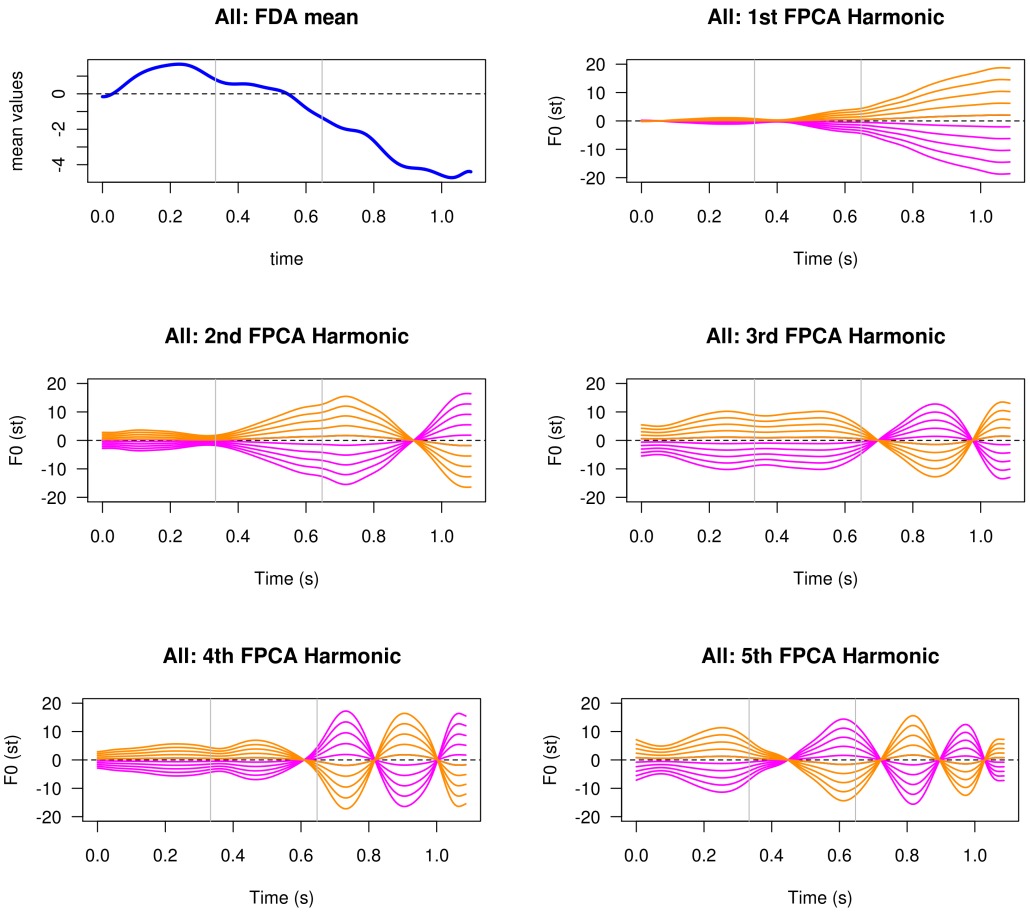


Figure 5.20: FDA Mean and Harmonics: Time warped, whole utterance contour. Orange and magenta lines indicates positive and negative scores respectively. The grey lines indicate landmark position.

	FPC1	FPC2	FPC3	FPC4	FPC5
lw	0.783	0.133	0.043	0.022	0.007
warp	0.667	0.115	0.071	0.040	0.030

Table 5.3: FDA: Proportion of Variance accounted for the first five principal components.

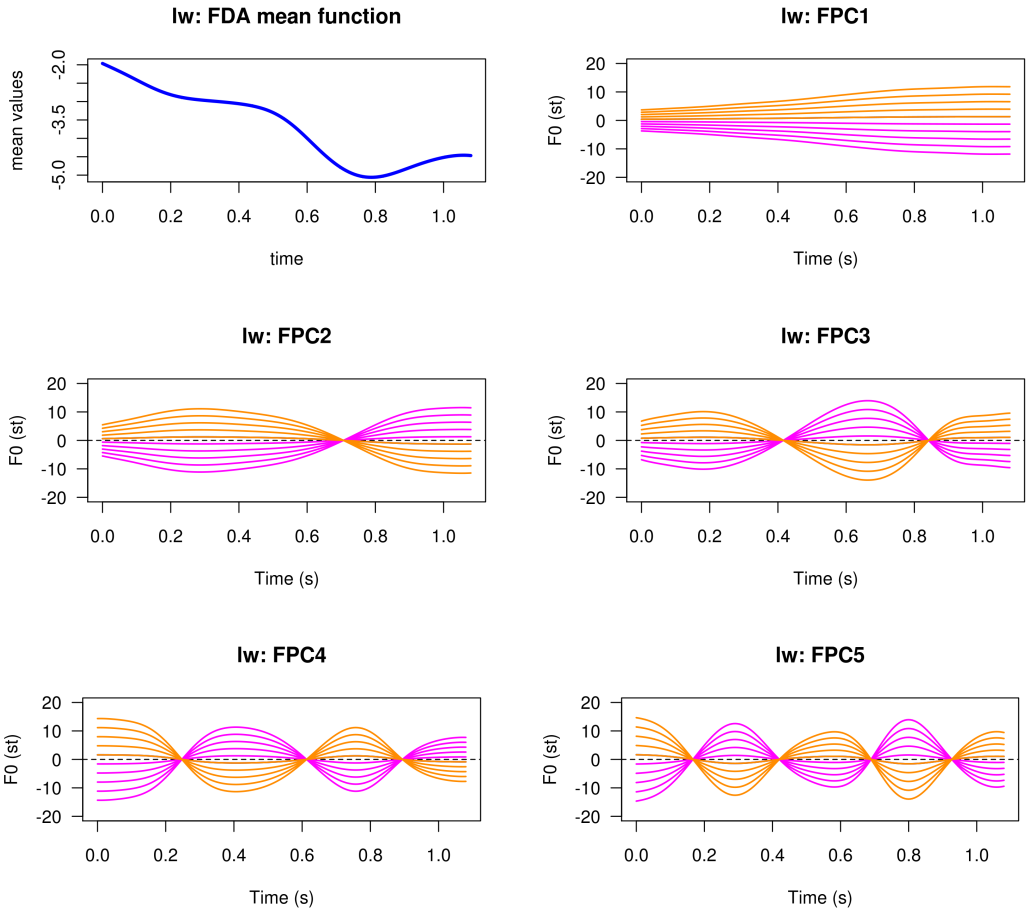
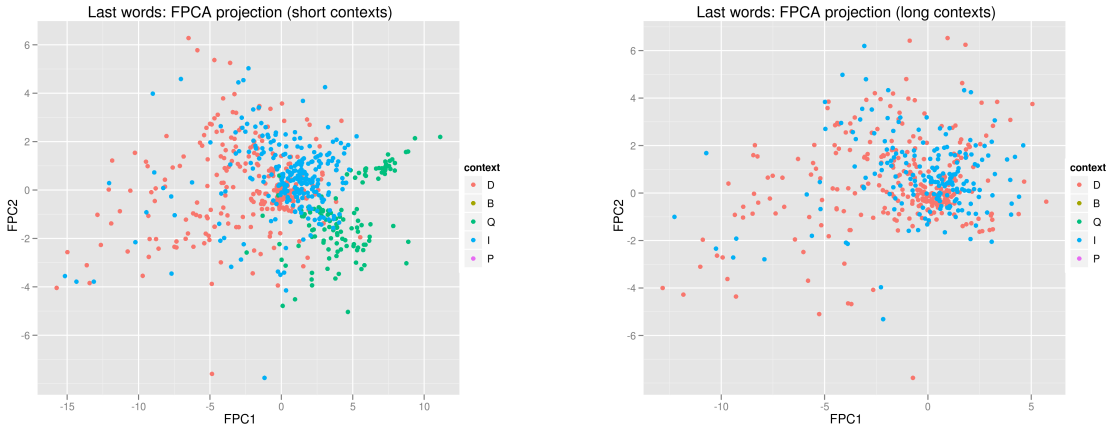


Figure 5.21: FDA Mean and Harmonics: Last words (fda.lw)



(a) Short contexts

(b) Long contexts

Figure 5.22: FDA projection: Last words (fda.lw) for short and long contexts (direct and indirect responses only) based on the short context production FDA.

harmonic suggests that a combination of the general height and tilt of the contour indicates the mode of greatest variation in the data. This accounts for 78.3% of the variation in the data (c.f. Table 5.3). The second harmonic gives us some notion of convexity/concavity at the utterance end (13.3% of the variation). Higher harmonics gives account for the finer variation in a similar way.

The distribution of scores for the first two harmonics (Figure 5.22a) shows clusters for the different response types which blend together towards the origin. From this we can see the differences between the three categories on the one projection. On the one hand, indirect and question responses generally have positive score for FPC1 (i.e. they are relatively rising) while direct responses are negative. On the other hand, the mass of indirect responses are higher in the second harmonic than the questions. This combination reflects the prevalence of the fall-rise shape at the end of the the indirect responses, and conversely scooping rises for the question type responses and steeper falls for the direct responses. So, while height and tilt capture a lot of what's going on, this type of decomposition still indicates that more fine grained modelling of the prosodic features, particularly convexity, is helpful for understanding what is going on in the data.

Figure 5.22b shows the long context direct and indirect responses projected onto the functional space from the FPCA on the short context productions. In this case, we still see that the the locus of the indirect responses is still in the upper right quadrant, while the direct responses spread onto the negative x-axis. However, the distinction between the response types is not so clear. Moreover, we would like to see whether looking at these sorts of shape features improves on using using the 'normal' prosodic aggregate features in separating out the categories. The next section investigates this

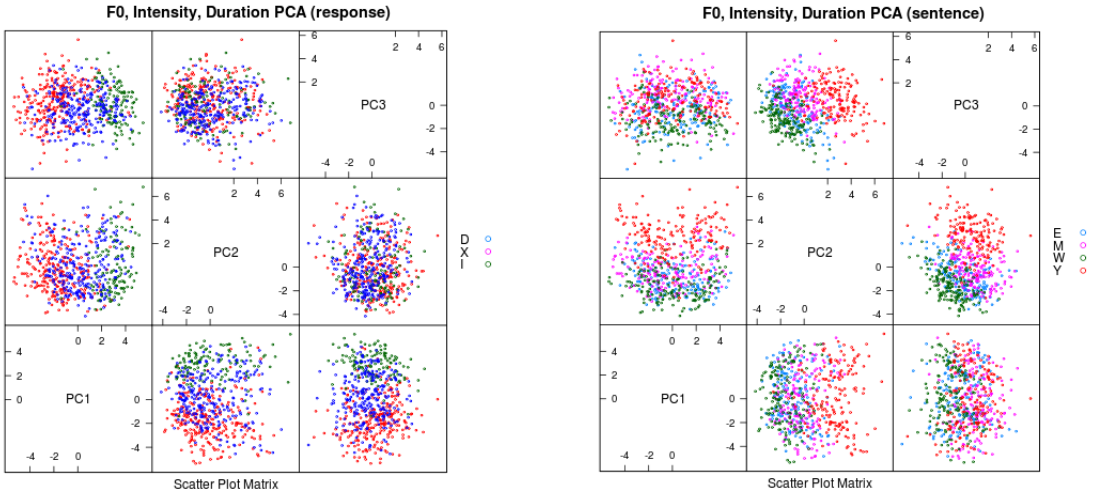
	Error	95 CI	F-measure	95 CI
Base	60.22		0.2303	
DTree (f)	30.38	(27.69, 32.32)	0.6944	(0.6747, 0.7204)
DTree (i)	44.43	(41.78, 46.87)	0.5501	(0.5258, 0.5759)
DTree (fd)	29.03	(26.60, 31.93)	0.7086	(0.6799, 0.7333)
DTree (ifd)	30.96	(28.56, 33.42)	0.6905	(0.6659, 0.7144)
DTree (f.pca)	30.56	(28.64, 32.95)	0.6942	(0.6702, 0.7137)
DTree (fd.pca)	30.50	(28.40, 32.56)	0.6950	(0.6746, 0.7163)
DTree (ifd.pca)	33.89	(31.46, 36.46)	0.6592	(0.6328, 0.6842)
SVM (f)	27.92	(26.92, 29.11)	0.7222	(0.7100, 0.7321)
SVM (i)	39.09	(37.56, 40.61)	0.6033	(0.5880, 0.6177)
SVM (fd)	**24.69**	(23.55, 25.82)	0.7523	(0.7411, 0.7638)
SVM (ifd)	25.39	(24.17, 27.07)	0.7469	(0.7301, 0.7593)
SVM (f.pca)	26.89	(25.98, 27.78)	0.7293	(0.7199, 0.7390)
SVM (fd.pca)	25.51	(24.33, 26.76)	0.7440	(0.7314, 0.7559)
SVM (ifd.pca)	26.18	(25.04, 27.07)	0.7374	(0.7287, 0.7489)

Table 5.4: Last Word Aggregate Features: Classification Results in terms of error rate and weighed F_1-score. f=F_0 features, d=duration, i=intensity features.

via some classification experiments.

5.4.6 Separability of Prosodic features

Inspection of contour and feature means in the previous section indicated that the productions varied based on the response type: direct, indirect and question. However, we would like to get a more concrete idea of how the prosody of the different response types varies and what we need to measure to capture this. Moreover, as we saw for the *really* data in Chapter 3, statistically significant differences between means do not always translate into realistically distinct distributions. So, we would like to get a better idea of how distinct the intonation of the different response types actually is. In this section we investigate these issues of variability and separability via through a classification experiments. More specifically, we compare the performance of classifiers trained on different feature sets drawn from the short context productions though which we investigate the separability of the training data, as well as the robustness of the classifiers with respect to unseen data from the long context productions.



(a) Response Type (b) Sentence Stimuli

Figure 5.23: First Three Principal Components

Aggregate Features

The first series of classifiers we consider are compare classification into direct (D), indirect (I), and question (Q) response classes based on the aggregate F0 (f), intensity (i) and duration (d) features, PCA based transformations of those feature vectors (e.g. ifd.pca).[9] Decision tree classifiers (J48) and SVMs were trained for each of the feature sets on the short context data using the R interfaces to the Weka machine learning toolkit (RWeka) and libSVM (e1071).

Table 5.4 shows mean error rates and F-scores for ten fold cross-validation over 100 randomized runs. The table also reports 95% intervals over these runs to give an idea of how the results change with different items in each fold. The classification results show that the SVMs outperformed the decision trees. However, the relative difference between feature sets has the same pattern for both types of classifier: the F0 and duration set (fd) performs the best. The addition of intensity features to the training data does not really add anything. As expected the classifier based on only intensity features performed the worst of the lot.

The classifiers based on the PCA transformed features don't improve classification performance. However, looking at the transformation itself sheds light on why the intensity features don't really seem to help above the F0 and duration features. Fig-

[9]Aggregate features for F0 and intensity were calculated over the last word in the utterance: mean, standard deviation, slope, jitter, maximum, minimum, absolute range, absolute and relative times of maximum value (minimum similarly). Duration is z-score normalized duration for that word. See Section 3.3.2.

	Error	95 CI	F-measure	95 CI
Base	59.93			
DTree (f.lg.last)	29.41	(27.62, 31.85)	0.7040	(0.6797, 0.7222)
DTree (f.lg.lw)	29.33	(27.30, 31.69)	0.7049	(0.6817, 0.7246)
DTree (fda.last)	31.34	(29.81, 33.34)	0.6868	(0.6667, 0.7020)
DTree (fda.lw)	31.42	(29.18, 33.57)	0.6860	(0.6639, 0.7081)
DTree (fda.warp)	31.80	(29.50, 33.80)	0.6823	(0.6619, 0.7053)
SVM (f.lg.last)	26.52	(25.43, 27.54)	0.7324	(0.7226, 0.7432)
SVM (f.lg.lw)	26.59	(25.27, 28.02)	0.7316	(0.7173, 0.7452)
SVM (fda.last)	25.84	(24.73, 27.00)	0.7401	(0.7283, 0.7512)
SVM (fda.lw)	24.46	(23.55, 25.67)	0.7556	(0.7435, 0.7649)
SVM (fda.warp)	**22.61**	(21.36, 23.87)	0.7745	(0.7617, 0.7869)

Table 5.5: Legendre polynomial and FPCA (fda) based coefficients: Classification Results

ures 5.23a and 5.23b show the projection of the data onto the first three dimensions for the PCA based on intensity, F0 and duration features. The first component (PC1) is dominated by F0 mean and slope, and we see a reasonable separation on this dimension for response type. The second principal component is dominated by intensity features, particularly the range. The second plot shows this is mainly differentiating the different sentences, rather than the discourse features.

Function based features

The cross-validation classification results for classifiers based Legendre polynomial (f.lg) and FPCA based coefficients (fda, cf. Section 5.4.5) are shown in Table 5.5. We look at coefficients derived over the last word (.last) and the segment from the last contrast position (.lw). For the FPCA set, we also looked at features derived from the whole utterance contour (fda.warp). The results are on par with the best classifiers produced using the aggregate features in Table 5.4. This reduces the number of parameters to just 5.

The small extension of the contour domain before the last word did not improve matters. However, we find that the SVM trained on the time-warped utterance contour (fda.warp) provided somewhat better results than we have seen for any of the other classifiers. So, it seems that information earlier in the contour does make a difference, although this is not something that can be used by the corresponding decision tree. Note, though, that at this point this sort of warping is fine because the utterance lengths don't differ too greatly. However, this does not seem tenable for classifying new longer utterances, even if we have the IS structure. So, this may be

be a case of overfitting.

Summary

Our primary concern is the relationship between form to function. So, the point of these experiments was to first to determine if the production of the different response types was different and if so, how different were they. Answering the second question entail choosing a criteria of difference which in turn depends on what we choose to measure, i.e. what features we use. We approached this problem by looking at how well different classifiers performed when trained using various feature sets. As an answer to the first question we find that the three response types produced in the short contexts (direct, indirect and question) differed in F0 contour shape in ways that could be separated by classifiers with an error rates of around 25%. So, unlike the dialogue act classification of *really* we looked at in previous chapters, the different response types looked at in the current study do appear to consistently induce different intonational forms.

For all the decision trees, highest level split is 'about' the end height of the contour. SVM projection too shows that the data basically divides on F0 height. Results were comparable using the parameterization produced by either FPCA or Legendre Polynomial based features. These in turn were comparable with the results for the classifier built on aggregate F0 and duration features of the utterance final word. However, we found that the best performing classifier overall was the one built on FPCA scores derived from the whole utterance contour, time warped with respect to two landmarks: end of subject/topic and the start of the last potential point of contrast. This suggests that the shape of things before the end also helps to determine what sort of discourse move is being played. However, it is not clear how robust the time warping is when applied to longer utterances. In the next section we will see how well this summarization of the data extends to new data.

5.4.7 Predictive Ability

The previous set of experiments focused on the separability on the short context productions. Given that the contexts were, indeed, extremely short and the fact that there was no continuity between statement/response pairs, we expect the responses to be more canonical and uniform in their manifestation. As such, we would like to know how well those classification/separability results reported above extend to the long context productions. In the previous section we didn't find much of a difference in error rate looking at classifiers built on aggregate and function based features. So, one goal for the current analysis is to find out which feature set describes the pattern

	Target-DT	Verum-DT	Target-SVM	Verum-SVM
base (I)	33.33	50.00	33.33	50.00
f	28.49	33.06	21.03	31.51
fd	27.22	42.34	37.84	48.09
ifd	32.46	35.61	43.18	46.01
f.pca	27.28	37.91	18.92	33.11
fd.pca	20.87	32.64	20.97	34.14
ifd.pca	33.74	38.33	18.85	40.25
f.lg.last	32.69	40.93	15.83	34.63
f.lg.lw	32.69	29.99	19.99	**23.28**
fda.last	**16.77**	**26.78**	**14.72**	27.94
fda.lw	37.74	30.82	19.96	*23.58*
fda.warp	37.77	43.02	35.85	37.61

Table 5.6: Classification of long context productions: error rate for Decision Trees and SVMs trained on various feature sets from the short context productions.

most robustly. Another is to gain further insight on where the locus of prosodic variation for the different response types.

Performance on the long context counterparts of the short context targets (EMWY) will give us an indication of how much the context and task difference effects the production of these contours, given that their basic structure is in a way expected by the classifier.[10] Eye-balling the data in Section 5.4.4, it seemed like the shape of target productions in the longer contexts matches what we have come to expect from the short context productions given the response type, although the gestures were somewhat bigger when the speakers were requested to sound more involved in the dialogue. In this vein, we saw a greater prevalence of the contradiction contour for the direct contradiction target utterance (for sentence Y). However, this variation is obscured if we only look at mean contours and aggregate features.

So, we would expect performance here to be quite good for E, M and W, if our representation do indeed capture the distinguishing prosodic points. On the other hand, we would expect a significant number of Y productions to be classified as Questions (because of the contradiction contour low rise). The long context productions also include several other direct and indirect responses involving verum focus with varying lengths and points of contrast. These productions present actual new data with which to test the classifiers for which the IS focus type is unambiguous (i.e. verum/polarity).

[10] In terms of timing, segmental effects etc. This is our control group

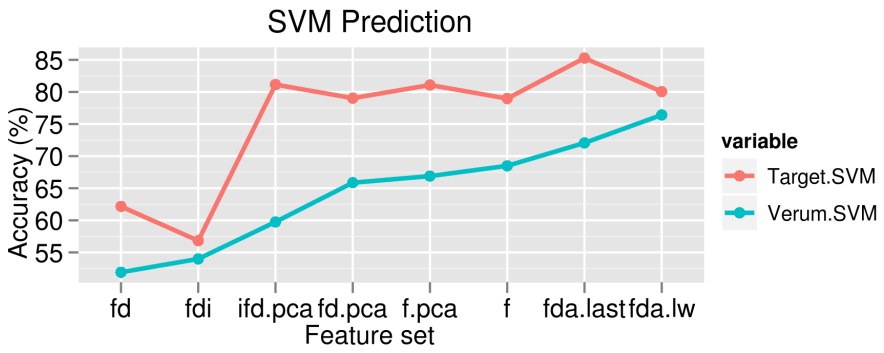


Figure 5.24: Visualizing the difference in prediction accuracy between seen and unseen data.

Table 5.6 shows shows how well the short context trained classifiers predict the response type of utterances from the long context recordings in terms of error rate. The table shows the prediction accuracy for the E, M, W sentences (*Target-SVM*, we treat the Y utterances separately) which appeared in both long and short contexts (we leave the Y case for later), as well as 6 other direct and indirect responses involving verum focus, unseen with respect to the training data (*Verum-SVM*). The FPCA based classifiers perform the best for both these groups (with the Legendre polynomial based classifiers performing at a similar level. Accuracy with respect to the Verum group is around the same level as the cross-validation results for the short context data (72-76%), while the Target group scores several points higher (e.g. 85% for fda.last). So, it seems that the shapes of these was consistent moving from the shorter contexts to the longer contexts, while the actual production of the response types was more distinct, probably due to increased speaker involvement.

Earlier we saw that the raw aggregate feature set of F0 and duration produced more or less as good a classifier as the functional feature sets for the short context data. This was not retained for the raw features here, though we do get comparable results for the PCA transformed data for the SVMs. While the FPCA based classifiers perform better with the new data than those based on aggregates, it is worth noting that the F0 only classifier (f) is not that far behind. The ifd.pca does not do so well here, which suggests that the classifier was overfitted in the separability experiments. This suggests that F0 characteristics are generally more robust indicators of response types than duration and intensity. We also get somewhat better performance when we take some structure into account than when we just take the last word for the unseen data with the SVMs but not necessarily for the decision trees. The improvement in accuracy for the functional features in the SVMs is visualized in Figure 5.24.

The results for the FDA warped contour set are much worse than the previous

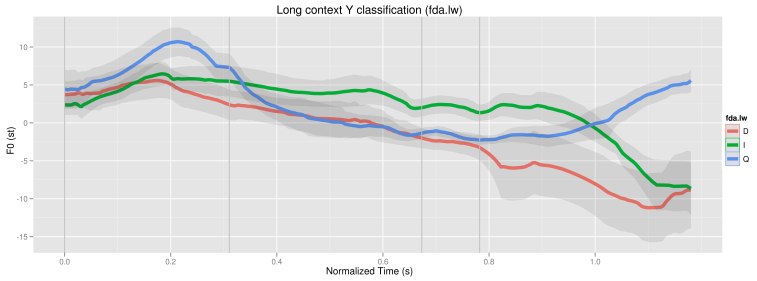


Figure 5.25: FDA Y predictions from the fda.lw classifier. Vertical lines indicate word boundaries, although this classifier only considers the data from the last word 'birthday'.

cross-validation results. This seems to be at least in part due to artifacts introduced in the landmark registration process. So, for now our best bet appears to be to focus on the utterance end and we leave the investigation of how to incorporate contour information from other parts of the utterance for future work.

5.4.8 Using the Classifiers to Explore Variation in the Data

While the classification experiments indicate that there are substantial prosodic differences between the elicited responses types, we can clearly see a lot of overlap between classes. In this section, we will see how the classifiers and the errors they make can be used explore how the categories vary. More specifically, we look at the intonational forms associated with broad focus contradictions (sentence Y), and pair-list answers to wh-questions. This is done, again, to probe at the question of how strong a connect there is between posited discourse/IS categories and intonational forms.

Direct Contradictions

Out of all the target sentences, the direct contradiction use of broad focus sentence Y seemed to have the greatest variation in form, regardless of the basis of the feature representation. As noted previously, this utterance was with the contradiction contour (Liberman and Sag, 1974) in a substantial number of cases. Figure 5.25 shows the mean contours and confidence intervals grouped by fda.lw classification. This classifier split the Y productions into three evenly sized groups (D=11, I=10, Q=11). Figure 5.26 shows the data projected onto the FPCA space. We see that the Q labelled contours, indeed, have the shape and distribution of the contradiction contour. The difference between the I and D labelled contours is that the former presents something of a hat contour, while the latter looks more like a linear fall.

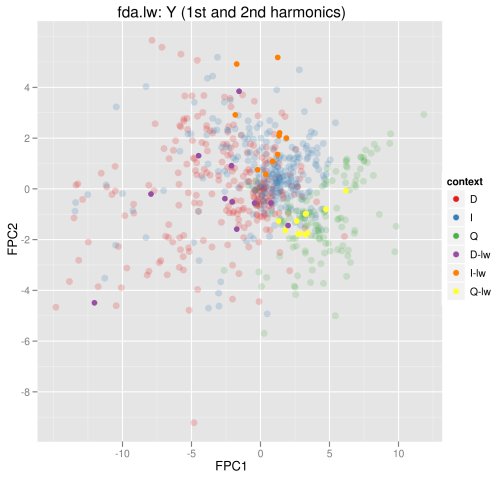


Figure 5.26: Projection of Y data for fda.lw

In this way, we can see how classifiers can be used to investigate what contours are available for a particular response type.

The presence of the conventionalized contradiction contour is quite reasonable since, after all, the production is a contradiction. However, in this case, the contradiction is also a correction about the properties of Mary. That is, a correction of VP: it is a property of *Mary* that she *remembered your birthday*. In this case, we would expect prominence on the topic, 'Mary' as well as the corrected item, the VP, inducing prominence on 'birthday'. So, having the type of contour we have been associating with indirect responses is quite plausible. The main motivator of prominence placement in both cases is the marking of contrast. The contour shape doesn't seem to indicate whether this contrast is corrective. However, we still might expect corrections to elicit more potential contrasts than affirmations in general.

Pair-list responses

The indirect responses were elicited in contexts that broadly match those which Büring (2003) associates with contrastive topics. That is, responses that don't directly resolve the current question under discussion. We would still like to be able to compare type of contour produced in that sort of scenario with the partial answers to wh-questions that Büring focuses on. More specifically, we would like to see if IS focus units can also bear a fall-rise type accent.

We expect pair-list type answers to wh-questions to have two contrasts: one associated with each of the wh-words. The wh-words in turn make up part of the IS focus or the ground. For sentences LB and MB, the subjects of answers are given

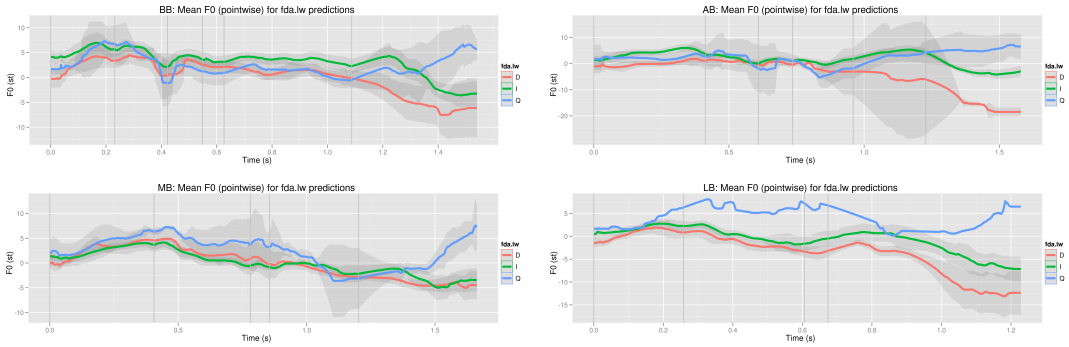


Figure 5.27: FDA pair-list predictions from the fda.lw classifier

from the preceding question, so these should be part of the IS ground. For sentences BB and AB, the pair-list wh-question is implicit. In both cases, the complete answer to the question has not been given after the second answer pair, so this fits Büring's contrastive topic context.

(73) B: So, what did the others do?
A: *Marianne* provided the BUILDING SCHEMA (MB)
A: *Lenny* provided THE PASSWORD (LB)
A: But we don't know who else is involved...

(74) A: No-one got me so much as a card
(i) Who got Gran what?
(ii) What did Bill get Gran? What did Annemarie get Gran?
B: *Bill* got you a CASHMERE SHAWL (BB)
B: *Annemarie* got you THAT TREE FERN (AB)
B: And see! Mary remembered your birthday

Both topic and focus are contrastive in the sense of invoking alternatives (i.e. F-marked). This gives them a similar IS to our verum indirect responses which also seem to involve a focus plus a contrast in the ground. However, in this case the order of the IS units is reversed since the pair-list answers have IS focus on the object.

The classification results show the pair-list responses do have a similar shape to the verum indirect responses in many cases. The mean pointwise contours for the different predicted classes (fda.lw) are shown in Figure 5.27. The sentence AB has the most identifiable fall-rise accent on the last noun phrase ('*tree* fern') while LB is less so. We can attribute the lesser presence of the terminal rises for the latter to the discourse structure – it's the effective list end, while AB is not. This trend is visible in Figure 5.28 which shows the projection of this data onto the the first two harmonics for the last words FPCA (fda.lw). Several of the LB productions have a

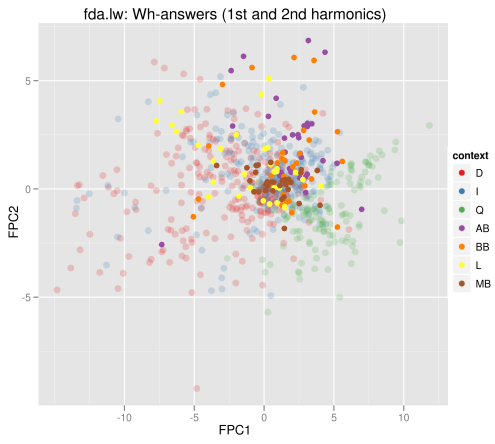


Figure 5.28: Projection of pair-list data for fda.lw

negative score for the first harmonic (i.e. downward tilt), while the rest are basically situated in the same position as the cluster for indirect responses. So it seems that in the same way that the production study elicited contrastive topics that didn't bear a terminal rises, and is not too hard to elicit a fall-rise accent on an IS focus. This casts doubt on the idea that particular accent shapes mark specific IS categories. In particular, we don't want to attribute difference in IS categories to the presence of phrase terminal rises.

5.5 General Discussion and Summary

We started off from the question of whether different response types have distinctive prosody, and how this is mediated by information structure. To investigate this we looked at verum focus declaratives. The reason for looking at this type of construction was that it has a easily identifiable IS focus which acts on propositions. We also had the intuition that indirect responses with verum focus had to have a fall-rise type accent in the post-IS-focus region (see Section 5.3.1). This was consistent with the hypothesis that terminal rises indicate that the discourse has not come to an appropriate stopping point, since indirect responses technically leave the current QUD unresolved. Conversely, we hypothesized that direct responses should exhibit deaccenting through the post-'did' region, ending in a fall. This deaccenting is what we would expect if VF in these situations was really narrow focus on polarity. Since VF draws the focus prominence away from the default utterance end position, post-focus prominences should mark units in the IS ground.

Generally speaking, these intuitions were confirmed in the production data. The production study showed systematic prosodic differences based on whether a declar-

ative was a direct or indirect response. These categories were broadly characterized by fall and fall-rise terminal contours respectively. If we were just to look at the relationship between prosody and their sentence type (declarative), illocutionary force (assertion), or the dialogue acts categories (statement, inform), we would not be able to see the patterns. This highlights the fact that to get coherent predictions about prosodic forms, we need to take the surrounding discourse structure into account.

Looking at this structure sheds light on why indirect responses are characterized by an IS unit having an unexpected level of prominence. In the verum examples, this is a second accent after the IS focus 'did'. For the broad focus utterances, we see a bigger gesture on the metrical main stress position. This extra prominence seems to mark a contrastive element which in turn evokes alternatives. For example, so the verum response in the following provides a partial answer for the question 'Who brought what?'.

(75) A: Nobody brought a dessert!
B: Emily did bring a meringue...

Answering the bigger question is a *strategy* for refuting the claim that 'Nobody brought a dessert'. In fact, Büring (2003) argues that fall-rise accents basically do this, i.e. signal a set of subquestions which are relevant for answering a question higher up in the discourse tree. The extra contrast gives a template for generating the strategy.

Büring, further argues that fall-rise accents mark IS units as being *contrastive topics*. However, experimental evidence suggests that such a strict mapping is not warranted. Like Calhoun (2007), our production data indicates that an actual fall-rise shape (as opposed to a fall), while common, is not necessary.[11] By evoking a strategy, via the extra contrast, indirect responses implicitly leave the current discourse in a open sort of state. While this is congruent with the idea that terminal rise signals non-finality, the rise isn't necessary to get this implication of openness. We can further note that some of the indirect responses do technically resolve the current QUD in an entailment sense while still eliciting a fall-rise accent, as in (75).

So, the presence of the fall-rise seems more to do with the QAC mismatch than whether the actual question under discussion gets resolved in the response in a possible worlds selecting sense. In general, intonational units don't act like semantic operators: a fall-rise accent can appear on units other than contrastive topics and a contrastive topic can be produced without the tail rise. So, it seems that to really investigate the

[11]The presence of the peak in close proximity to the boundary can be enough to make the contour sound rising (c.f. Studdert-Kennedy and Hadding (1973)). We'll come back the question of how rises in these situations are perceived in the next chapter.

prosody-meaning map, we need a notion of dialogue move that reflects not just the illocutionary force of an utterance but also how it fits into the dialogue structure, i.e. response types.

From the agreement/contradiction point of view we see very little difference between productions in terms of terminal rises (although obviously the question class was well defined in this way). However, we did find direct contradictions to fall less steeply than direct agreements on average. In fact, some contradictions were produced with low rises, besides those produced with the full contradiction contour. The presence of these continuation like rises is understandable given that contradictions leave the discourse in a controversial state which is, again, not a good stopping point. We see a more striking differences between direct agreements and contradictions on the inserted 'did' for the verum cases, with contradictions having generally bigger and more delayed peaks. In the contradiction cases, the polarity of the proffered content is clearly under dispute, so, this goes against Steedman's (2007) hypothesis that H*/L* distinction in accent shape primarily distinguishes agreed/non-agreed upon informational units respectively. Instead, it seems more likely that corrective prominence are produced with more effort because correction is a more marked/salient sort of move in general. From this point of view the concavity of the accent associated with declarative questions seems to mark a potential correction/check point rather than something that is actually under dispute.[12] This would further suggest that H* is simply a default prominence marker, while L* (convex) accents convey some discourse information by virtue of being a marked option.

In Sections 5.4.6 and 5.4.7 we used machine learning tools to investigate how we could describe these features in terms of parameters. In particular, we wanted to see if those feature distributions formed distinct clusters so we could get some idea of how the data varied. We found that using orthogonal polynomial decompositions to be an effective way to describe the data with a reasonably low number of parameters. Besides the Legendre polynomial coefficients we have been working with in previous chapters, we looked at decomposition based on functional principal components analysis. This gave us a more concrete way looking at the modes of variation within the data set. While the aggregate features over the last word in the utterance had comparable cross-validation results to the functional feature sets, they did not generalize as well to the new data. This highlights the utility of capturing information about the convexity of a prosodic event, rather than just means and single points of interest, like the position of the maxima.

Overall, the basic differences observed from auditory impressions, direct inspection of the F0 contours, and the distribution of features – aggregate and functional – point

[12]For example, a unit that is improbable but not unbelievable: e.g. 'You got the job? That's great!'

to a description of the data by nuclear tone shape. On average, direct responses fall from the nuclear accent, indirect responses have more of a fall-rise, while the declarative questions have scooped rise. So, it seems that the falling off-ramp of the nuclear accent is what the assertions have in common, rather than a low boundary tone (c.f. Bartels (1999)). However, when we look at the longer context data we find more direct contradictions ending in rises. The presence of such a rise is compatible with the stopping point analysis of terminal rises, since contradictions leave the discourse in an unstable state. However, it is potentially problematic for an analysis of rising *accents* as signalling a check (non-assertive) move. We can overcome this problem by again noting that the CC contour is conventionalized as a whole utterance contour marking a contradicting assertion with a high initial and ending pitch. In fact, the terminal rise shape differs from that of DQs (c.f. Section 5.4.7) – it seems that the rise has to be bigger/more scooped to get the DQ interpretation. However, this was not distinguished in by our classification scheme.

What about rises?

So, what does this tell us about the the role of rises in dialogue? The picture we get from our analysis so far, is that if there is no QAC mismatch and the nuclear accent shape is concave/peaked then a terminal rise will be interpreted as the classic continuation, i.e. that there is something more the speaker wants to say. If, on the other hand, there is a QAC mismatch here, (e.g. narrow focus/extra prominence, no explicit wh-question), the move will be interpreted as indirect and strategy evoking (possibly witness/counterexample) whether or not there is a terminal rise. Because an indirect, non-QAC conforming, response was used, we get the signal that the QUD is not officially resolved. So, there are two paths to that 'rising feeling' of open-endedness. One is to have an actual rise. The second is to invoke some non-canonical move that gives the same implication. Our production data clearly shows that you can have both, so we would like to know if the presence of a rise is additive. If we have a parallel to the case of the cue word *really* discussed in previous chapters then we would not expect an additive interpretation. However, we would expect speaker involvement to increase as the gestures become more effortful. The next chapter investigates the interpretation of sentential prosody as with respect to this sort of data.

5.6 Conclusion

The main goal of this chapter was to show that we get a better understanding of the prosody-meaning map if we look at categories which more directly reflect dialogue

structure and expectations. The dominant factor determining these expectations appears to be question answer congruence, which is reflected in the information structure of the response. In the production study, we indeed found that prosody varied consistently by looking at response dimensions like whether an utterance directly or indirectly addresses the question under discussion. We confirmed the idea that indirect responses represent discourse contexts that can (and often do) elicit a terminal rise as part of a fall-rise accent in the IS ground. This fits with the fact that, given that a direct response was not given, a direct resolution of the current question under discussion was not possible or desirable. That is, these discourse configurations do not present good stopping points in the dialogue.

Machine learning experiments show that we can separate out the different response types based on aggregate statistics like mean F0 to a fair degree. However, the differences in the prosody are more succinctly categorized by the overall contour shape from the last accent to the utterance end. We quantified these differences using functional data analysis techniques. This allowed us to quantify contour shape over a continuous space with minimal manual annotation effort: we only have to specify the domain, rather than a series of inflection points which may not be consistently applicable in every case (e.g. pitch 'elbows' are not always identifiable for 'weak' accents Calhoun (2007)).

Looking at the distributions of such features highlights the fact that while indirect responses are often produced with a distinct fall-rise accent, they do not always exhibit the rise part. The next chapter takes a closer look at what this rise adds to the interpretation of such dialogue moves from perceptual and theoretical perspectives.

Chapter 6

Interpreting the Rise in a Fall-Rise

6.1 Introduction

One of the main results from the previous chapter was that indirect statement responses were, on average, produced with a fall-rise nuclear tune. However, while this response type more or less always elicited a prominence in the IS ground, the low rise at the utterance boundary was not always present. In this chapter we look at how the presence of such a rise in this sort of discourse configuration affects an utterance's interpretation. The cue word data from previous chapters indicates that rises don't mark *specific* dialogue acts or propositional attitudes. Instead, the data fit with the idea that rises signal discourse non-finality. The use of an indirect response implicitly suggests that the discourse has not come to a stopping point (otherwise the speaker would have resolved the current question under discussion directly). This suggests why rises are found in these sorts of configuration. It is not clear however, if they add anything more.

More generally, we would like to know whether rises should be treated as signaling something independent of the shape of the accent it is attached to. That is, should the contributions of pitch accents and boundary tones be treated as compositional. As we have seen, the former approach is espoused in influential works like Pierrehumbert and Hirschberg (1990) and Steedman (2000). However, several studies that have their roots in that work have taken a more holistic approach in studying intonational meaning. For example, Gunlogson (2003) and Gunlogson (2008) restrict the analysis of rising declaratives to those rising from the nuclear accent. Nilsenova (2006) similarly excludes continuation type rises from her examination of final rises as contributing an epistemic modal type operator. On the other side, Constant (2007) analyzes fall-rise accents as contributing a focus sensitive quantifier like *only*.

In either case, it is not clear how the types of rises and fall-rises dealt with in these theories relate to each other. Intermediate to these nuclear tone type approaches, Büring (2003) analyzes fall-rise accents as marking contrastive topics as a whole, but suggests that the locus of the focus/contrastive topic distinction should actually be the boundary tone. Wagner (2010) argues that fall-rises only produce the feeling of unresolvedness (associated with contrastive topics) when the rise is utterance final and so the 'scope' of the rise is sentential. However, how this fits into the larger intonational picture is not pursued.

In order to get a better idea of the contribution of the rise in an utterance final fall-rise, we look at how the interpretation of indirect responses changes with and without this rise. More specifically, we look at the perceived attitude of the responder towards what the other speaker has said and the state of the discourse. This is investigated through a perception experiment using resynthesized stimuli from the production experiment. As suggested above, the main goal of this experiment was to determine whether the rise has an additional effect beyond what is available from the discourse/IS structural configuration. We also investigate the relationship between the size of an intonational gesture and perceived speaker engagement.

We use the perceptual results to evaluate the plethora of analyses of the meaning of the fall-rise accent. These analyses come in two main threads: those for which FR labels an information unit, usually contrastive topic (Jackendoff, 1974, Steedman, 2000, Büring, 2003), and those for which it signals some sort of relationship to an alternative proposition (Ward and Hirschberg, 1985, Constant, 2007, Wagner, 2010). The common thread to all these analyses, however, is the observation that utterances with a fall-rise nuclear tune provoke a feeling of incompleteness, or something having been left unsaid. That is, fall-rise accents evoke unevaluated alternative propositions. The differences between these analyses stem from how the relationship between these alternatives and the discourse structure is modelled.

The data collected in the experiments lead us to the conclusion that the meaning associated with rising contours like FR is best analyzed by separating the contributions of the rise, the discourse structure and information packaging. From this point of view, fall-rise shaped accents don't have a one-to-one relationship with an IS category like contrastive topic. Instead, they are simply licit in some contexts that match the description of that particular IS unit. However, fall-rise accents can appear in other configurations, and conversely contrastive topic like configurations can appear without an FR accent. Nevertheless the core of what makes FR licit in the utterances we do see them is to be found in their relationship to discourse strategy: the notion that resolution of a task (i.e. resolving a question) may require breaking down the problem into subtasks (answering subquestions) (Roberts, 1996, Büring, 2003).

6.2 Perception Experiment: Fall-Rise, Fall-Fall

6.2.1 Goals and Expectations

In this experiment, the goal is to see whether a final rise attached to a falling accent adds different to what you get from a non-rise. To look at this we need some idea of what it might add. We would like to know if the intonational characteristics reflect the attitude of the speaker towards the proposition they are addressing (the current QUD), their own response, or the state of the discourse. This presents us with another opportunity to test the link between rises, uncertainty, as well as discourse unresolvedness and non-finality. These attitudinal aspects of an utterance are generally evoked in one way or another in analyses of the fall-rise accent (Ward and Hirschberg, 1985, Büring, 2003, Constant, 2007). The fact that the same sort of attitudes are invoked in the literature on nuclear rises suggests that the rises in fall-rise accents can be treated the same way as the rises on cue words.

A basic additive hypothesis would be that adding a rise increases the perception on some scale related to uncertainty. For statement-response type contexts we have been examining, this uncertainty can be directed at various parts of the dialogue.

(1) The speaker is uncertain about the evaluation of the utterance they are responding to. (evaluation uncertainty)

(2) The speaker is uncertain about whether their response resolves the salient QUD. (discourse uncertainty, non-finality)

The production data presented in the previous chapter already argue against the idea that fall-rises are in a 1-1 relationship with a single IS category like contrastive topic. In contrastive topic contexts, however, we might still expect that a rise increases the perception that a response only partially addresses the question at hand (i.e. non-finality). If we don't see a difference between falls and fall-rises in the same IS contexts, then we should be able to say with certainty that the rises do not differentiate these categories.

6.2.2 Data

The base stimuli used in this experiment were taken from a pilot run of the production experiment described in the previous chapter. The data was taken from the short statement-response portion of the experiment. The statements and responses were produced by a female and a male respectively. One production from the indirect

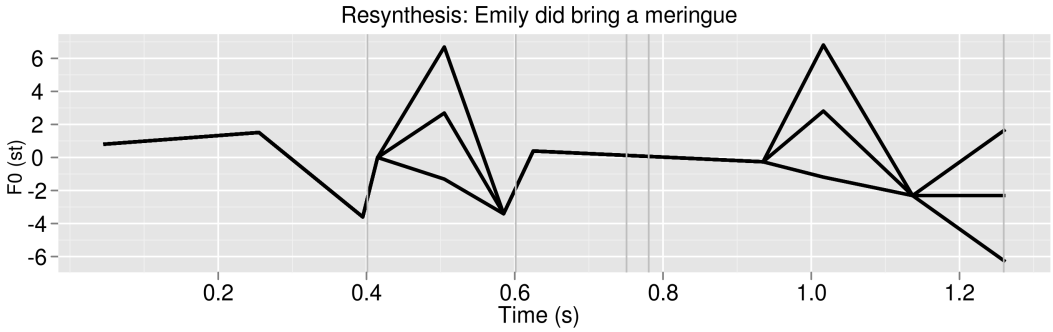


Figure 6.1: Stylized resynthesis for sentence E: Emily did bring a meringue (indirect).

contexts was selected for resynthesis. These were selected based on their exhibiting modal voice quality and a distinct fall-rise accent. One of the pair-list responses from the long context responses (BB=Bill got you a cashmere shawl) was similarly chosen for manipulation.

Pitch resynthesis was performed using PSOLA in Praat. Base stimuli were first linearly stylized (2 semitone threshold). The stylization identified two pitch peaks in the base stimuli. For the verum focus production the first peak fell on the auxiliary *did*, for the non-verum productions, the peak fell on subject. The resynthesis involved two manipulations. The two peaks were set at 3 levels:

(3) Peak manipulation:
   a. Level 3 = original (high)
   b. Level 2 = Level 3 - 4 st (medium)
   c. Level 1 = Level 3 - 8 st (low)

So, Level 1 basically gives a flattened out second peak. Note, intensity and duration features were unchanged, so even when the peak is flattened, there is still a signal that that word is metrically prominent. The utterance tail height also varied by 3 levels, where the tail was identified as the rising portion of the (second peak) fall rise accent. The start of the tail was the point of inflection (the pitch 'elbow').

(4) Tail manipulation:
   a. Level 2 = level with last 'elbow' (level)
   b. Level 3 = Level 2 + 4 st (rising)
   c. Level 1 = Level 2 - 4 st (falling)

Besides the resynthesized utterances, the stimulus set included direct statement responses and declarative questions in unmodified form. In this case, three productions

were selected for each of the target sentences (agreement, contradiction, question). Three such productions were also included from the *really* fillers: 'Anna (really) is (really) in Rio' (AR, AI), and 'The robbery (really) was (really) a decoy' (RR, RI). Two unmodified version of the pair-list answer (BB1, BB2) were also included. The expectation was that these direct responses would elicit scores on the extremes of the scales used in this experiment. The stimuli were matched with the appropriate preceding statement to produce the right discourse configuration from the production experiment. All together, 107 stimuli were presented in this experiment, summarized as follows:

- Indirect Resynthesized:
  {stim: E,M,W,Y} x I x {agree: A,C} x {peak:1,2,3} x {tail:1,2,3} (72)
- CT Resynthesized:
  BB1 x {peak:1,2,3} x {tail:1,2,3} (9)
- Calibration:
  {stim: E,M,W,Y} x D x {agree: A,C,Q} (12)
  {stim: AI,RI,AR,RR} x D x {agree: A,C,Q} (12)
  BB1 + BBC2 (2)

6.2.3 Method

In this experiment participants rated statement-response (A-B) pairs according to the following scales:

- CREDIBILITY: How **credible/correct** does what A said seem to B?
  - 1=not at all credible/correct,
  - 7=completely credible/correct
- NON-FINALITY: Does B seem to think that **more** things need to be considered to determine whether A's statement is correct?
  - 1=no, the correctness/incorrectness of A's statement has been resolved,
  - 7=yes, the correctness/incorrectness of A's statement is still open
- ENGAGEMENT: How **engaged** in the conversation does B sound?
  - 1=not at all engaged, he sounds bored/disinterested,
  - 7=completely engaged, he sounds interested/involved

- SURENESS How **sure** does B sound about whether A's statement is true or false?
  - 1=not at all sure,
  - 7=completely sure

Assuming that rises signal discourse non-finality, if intonation is additive this should elicit higher NON-FINALITY scores. Conversely, falls should produce lower NON-FINALITY ('more') scores. The CREDIBILITY scale indicates the polarity of B's evaluation of A's statement. This should be more or less independent of the prosodic form of the utterance, and instead be determined by the semantic content of the statement and response. The SURENESS scale reflects how certain the speaker is of their own evaluation. We would expect these scores to vary according to the 'off-ramp' shape of the final/nuclear accent. More specifically, if we link falls to assertions, we expect falls and fall-rises to sound sure but rises to sound unsure. Finally, from the cue word studies, we expect that as pitch gestures become more pronounced, perceived speaker engagement should increase. So, we should see higher ENGAGEMENT scores for higher peak levels and when the second pitch accent is attached to a rise.

The experiment was carried out online using Qualtrics survey software. The first three stimuli consisted of a direct agreement, contradiction and declarative question. The rest of the stimuli were randomized in 9 blocks to reduce runs of particular target type and contexts. Participants were allowed to complete the experiment at their own leisure. 15 speakers from the production experiment participated in this experiment (5 did not complete the whole experiment, but we still count the stimuli for which were rated).

6.2.4 Results

Distribution of ratings

Figure 6.2 shows the distribution of rating for the response types. The picture we get is that the different response types broadly map to different points on the attitudinal scales. As expected, the CREDIBILITY scores are inline with the polarity of the responses evaluation (i.e. agreement versus contradiction). The pair-list contradictions patterned with the direct contradictions. The indirect responses elicited less extreme values than the direct responses. This makes sense pragmatically given that the responder would have used a direct response if they were sure one way or another whether preceding statement was correct or not. This is reflected in the SURENESS and NON-FINALITY ratings. Here, in general the indirect response ratings were closer

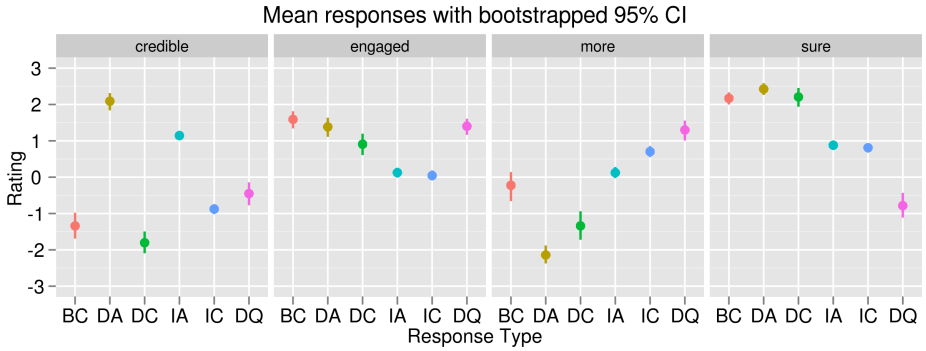


Figure 6.2: Centered scores by question.

to those for the declarative question. We also see that contradictions had higher ratings in this respect than their agreement counterparts, which is to be expected since by nature contradictions leave the discourse in an unresolved state.

PCA was applied to the ratings (as a 4 dimensional vector), to shed some light on how the scales relate to each other. Figure 6.3 shows a biplot, projecting the rating data and the original basis vectors on to the space spanned by the first two principal components. The graph also shows mean values for the different response types based on these PCA dimensions. From here, we see two main dimensions of variation with respect to the response type. The direct responses exhibit greater SURENESS about the responders evaluation of the previous utterance. On the other dimension, the contradictions signal that more needs to be said on the current QUD.

A non-nested multilevel linear model was fitted for each of scales to examine the size of the effects more closely. The model had the following form:

$$
\begin{aligned}
y_i &\sim N(\mu + \alpha^{\text{response}}_{j[i]} + \alpha^{\text{agree}}_{k[i]} + \alpha^{\text{rtype}}_{l[i]} + \alpha^{\text{stim}}_{m[i]} + \alpha^{\text{id}}_{n[i]}, \sigma^2_y), \text{ for } i = 1, \ldots, 1276 \\
\alpha^{\text{response}}_j &\sim N(0, \sigma^2_{\text{response}}) \text{ for } j = 1, \ldots, 4 \\
\alpha^{\text{agree}}_k &\sim N(0, \sigma^2_{\text{agree}}) \text{ for } k = 1, \ldots, 4 \\
\alpha^{\text{rtype}}_l &\sim N(0, \sigma^2_{\text{rtype}}) \text{ for } l = 1, \ldots, 6 \\
\alpha^{\text{stim}}_l &\sim N(0, \sigma^2_{\text{stim}}) \text{ for } m = 1, \ldots, 7 \\
\alpha^{\text{id}}_l &\sim N(0, \sigma^2_{\text{id}}) \text{ for } n = 1, \ldots, 15
\end{aligned}
$$

Note, the rtype factor is just the interaction term agree:response. The models were fit in R using the `lmer` package. The parameter estimates for response, agree and rtype factors are shown in Figure 6.4. These are basically inline with

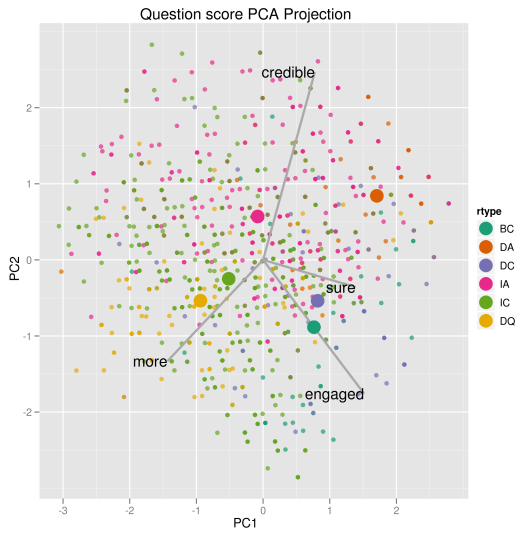
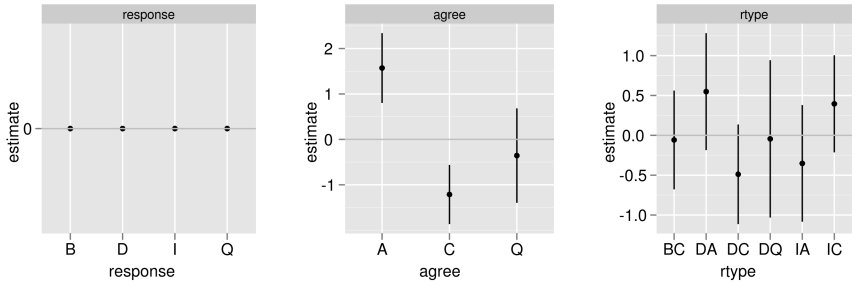


Figure 6.3: Different response types, PCA on scores.

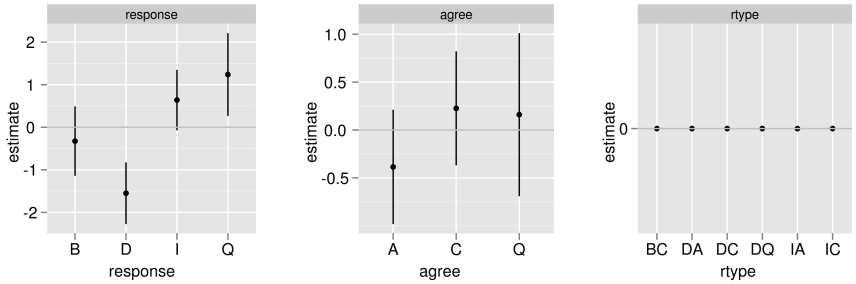
the observations above. We see significant effects on the credibility scale on the agreement/contradiction dimension. The trend for the interaction term is for direct responses to pull the ratings towards the extremes, while the indirect responses pull them towards the center. There are significant effects of response type for the SURENESS and MORE scale, with indirect responses reducing the ratings of sureness and increasing the perception of non-finality. We see something similar for the ENGAGEMENT scale. These estimates show how the different discourse factors lead to different perception about the responder's attitude. However, this doesn't take into account the prosodic variation introduced by the pitch resynthesis. To find out whether rises have an additive effect beyond these baselines, we need to look more closely at the indirect responses.

6.2.5 Resynthesized Indirect Responses

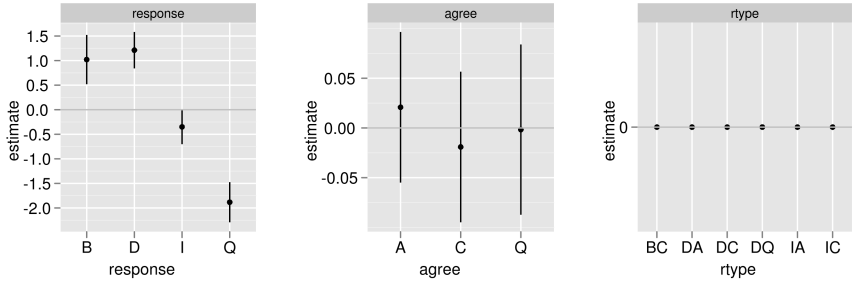
Figure 6.5 shows mean ratings for the indirect responses by peak and tail height manipulation. As expected, we see a positive relationship between peak height and level of engagement. In contrast, we don't see any other clear trends for the other scales. Based on the discussion above, we wouldn't expect pitch differences to have much of an effect on the credibility rating, since the the main signal for this scale is encoded in the actual semantic content of the response. However, under most accounts of fall-rise meaning we would have expected rises to increase the impression of non-finality scores, if not speaker sureness.



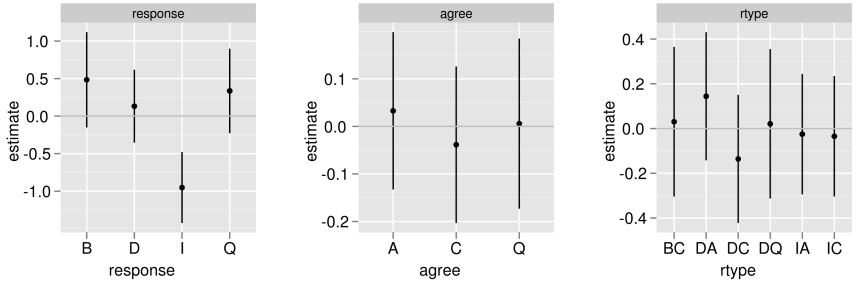
(a) Credibility



(b) Non-finality



(c) Sureness



(d) Engagement

Figure 6.4: Parameter Estimates.

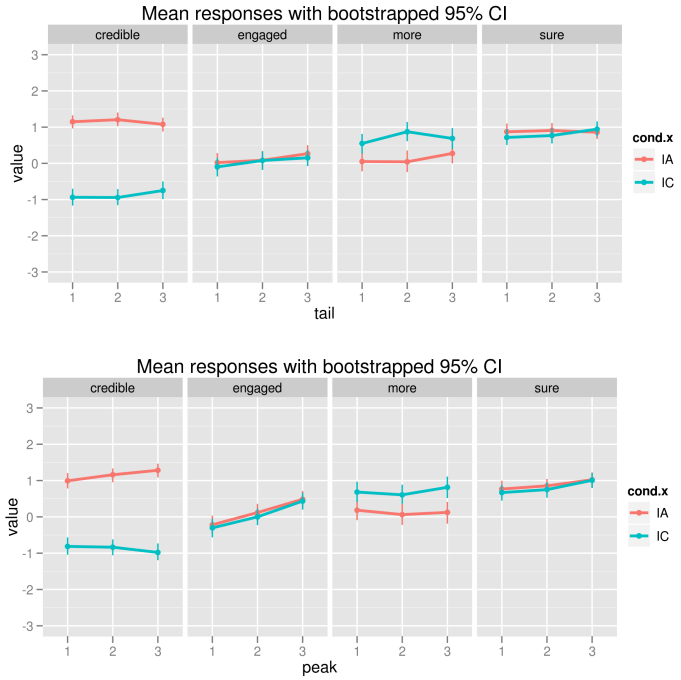
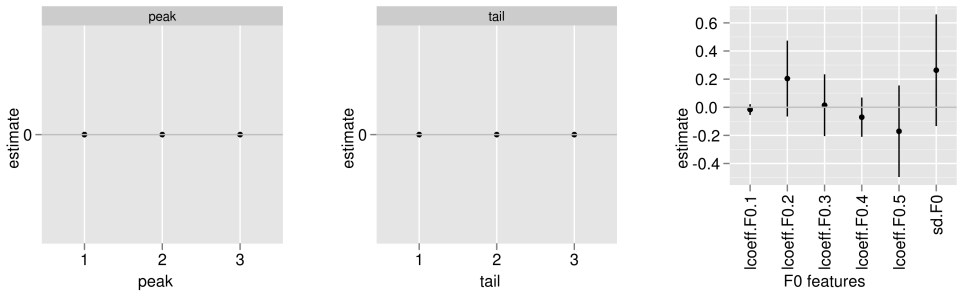
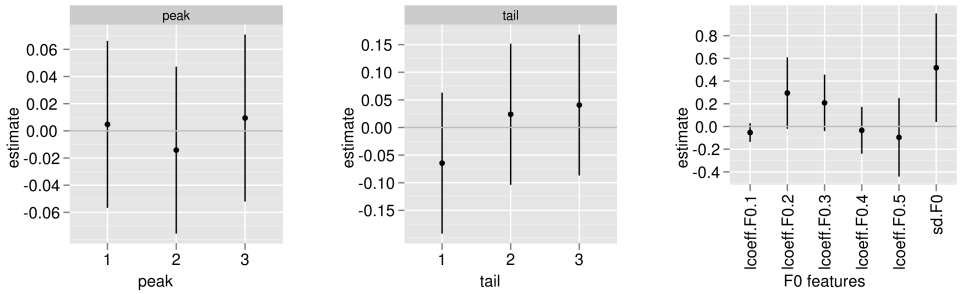


Figure 6.5: Indirect Responses: mean scores and 95% bootstrapped confidence intervals.



(a) Peak (b) Tail (c) F0 features

Figure 6.6: Credibility



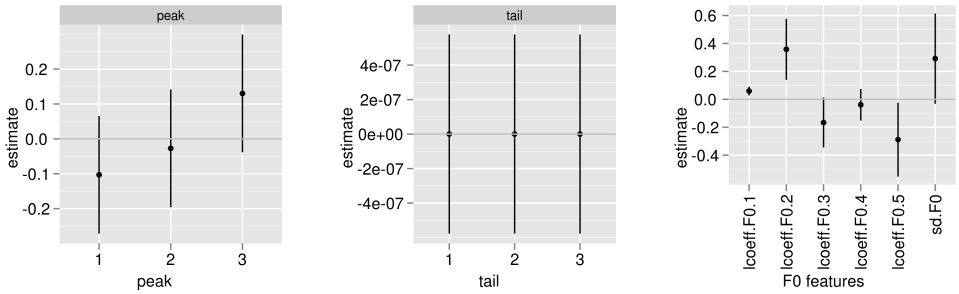
(a) Peak (b) Tail (c) F0 features

Figure 6.7: Non-finality

So, it seems at this point that a final rise in this discourse/prosodic context doesn't add anything on this dimension. However, it may be that we need to take the accent shape as a whole into account rather than just the rise characteristic. In the following, we investigate this by looking at looking at the effect of the pitch manipulation via pitch features derived from the signal, as well as the categorical representation from the resynthesis.

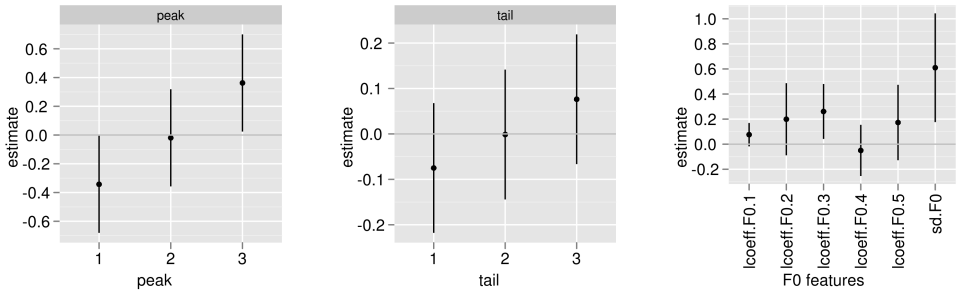
We first examine the effect of boundary pitch by including terms for peak, tail, and their interaction in the multilevel model similar to the one examined above above. For the indirect response data, however, we only need to consider the rtype contrast (IA versus IC). We also look at a model including Legendre coefficients fit from the peak and tail manipulated word as individual level (fixed) effects (cf. Section 5.4).

Credibility. Adding terms for peak and tail height and their interaction to the credibility model made no significant improvement in fit compared to the model without these terms (likelihood ratio test, $p > 0.05$). As for the general case above, we do see significant differences based on the agreement factor: being an agreement adds a point, while contradictions subtract one. Adding Legendre coefficients as individual



(a) Peak (b) Tail (c) F0 features

Figure 6.8: Sureness



(a) Peak (b) Tail (c) F0 features

Figure 6.9: Engagement

level effects does not significantly improve the model fit. Parameter estimates are shown in Figure 6.6. None of the estimates were significantly greater than zero. So, again it seems pretty clear that the credibility rating was not affected by the prosodic form of the response.

Non-finality. Similarly, the addition of the pitch features, categorical and acoustic, did not improve the model fit for the non-finality, MORE, scale. Out of the parameter estimates to do with pitch features (Figure 6.7), only the estimate for F0 standard deviation was significantly greater than zero. So, it seems that the pitch accent shape does not add much to the perception of non-finality. The estimated effects for tilt (lcoeff.F0.2) and convexity (lcoeff.F0.3) are on the positive side which suggests that rises increase the perception of non-finality if they do anything at all. However, the contribution of the rise is clearly overshadowed by what is being signalled by the discourse/IS configuration.

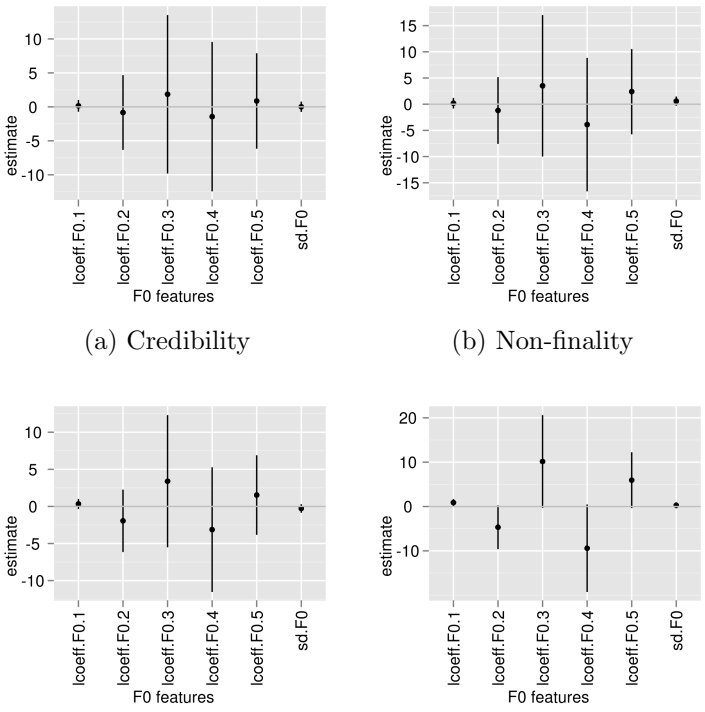
Sureness. Although the model including peak and tail manipulation did not produce a significantly better fit than the base model for sureness, we do see an improvement including the F0 shape features (LRT, $p < 0.001$). In particular, we see significant effects for height (lcoeff.F0.1) and tilt (lcoeff.F0.2). The sign of the convexity estimate (lcoeff.F0.3) is negative, which means that sureness ratings increased as the accent became more peaked, but less falling overall. This is inline with the idea that the more effortful the prosodic gestures, the more the speaker believes that the content of the utterance is important to the discourse.

Engagement. Unlike the other uncertainty type scales, We find significant effects on the engagement scale adding information about the accent shape the model. In Figure 6.9a we see that the ratings increase with peak height, and similarly for tail height (Figure 6.9b). The estimates for the F0 features (Figure 6.9c). This again supports the notion that increases in effort related features like pitch range (estimated with standard deviation) result a greater in the perception of speaker engagement. The link between final rises and engagement is reflected in the positive coefficient for the convexity measure (lcoeff.F0.3), and to a less extent the tilt (lcoeff.F0.2). In general, we see that more articulated F0 gestures heighten perception of engagement.

6.2.6 Pair-list Responses

The resynthesis set included a pair-list type response as a base utterance. This was from the longer dialogues in the production experiment and had the following form:

(5) A: Nobody got me so much as a card
B: Bill got you a cashmere shawl



(a) Credibility (b) Non-finality

(c) Sureness (d) Engagement

Figure 6.10: Pair-list (B) utterances

In the production experiment, B's response was immediately followed by another paired answer of the same form ('Annemarie got you that tree fern'), so in the recording the speaker was aware that the initial response was only a partial answer to the implicit question 'Who got granny what'.

We look at the effect of the pitch manipulation via the same type of model used for the indirect statement responses. In this case, however, we don't have any variation in terms of stimulus or actual response type as only one context was used. As was the case for the indirect responses, there is a significant improvement in model fit when including the F0 features for the engagement model. The parameter estimates are shown in Figure 6.10. None of the estimates quite achieves a significant difference from zero. Nevertheless, it seems that higher levels of engagement were associated with a more convex nuclear accent (lcoeff.F0.3), and negative tilt, i.e. more of a fall. Once again this supports the idea that bigger pitch gestures convey greater speaker involvement.

Overall, the prosodic changes did not have any discernible effect with respect to the credibility, sureness and non-finality ratings. The sureness ratings for this response type were high on average, while the non-finality scores were in the middle

of the scale (Figure 6.2). The high sureness ratings once again put something of a check on the idea of analyzing rises as saying something about the epistemic status of the utterance that carries them. For the latter result, we note that the distribution of non-finality scores for this response type was somewhat bimodal. This reflects the differing strategies of subjects, with some marking all these responses are final and others marking them as non-final. So, it seems that some subjects had strong expectations about this sort of discourse configuration which overrode the potential contribution of the terminal rise.

6.2.7 Discussion and Summary

In the production experiment we saw that these sorts of indirect and partial answer responses varied in terms of whether they elicited a fall or fall-rise type contour. That data argued against associating the fall-rise with a specific IS category like contrastive topic. On the flipside, the perception data indicates that having a fall-rise doesn't necessarily induce the type of feeling of discourse incompleteness or lack of resolution that has been associated with contrastive topics, above that contributed by non-rising accents (Büring, 2003, Wagner, 2010).

As it stands, the basic observations are as follows. As in Büring (2003) we find that question answer congruence determines to a large extent the prosodic form of an utterance in terms of prominence placement. It seems for statement/informing type moves, the default accent shape is concave, i.e. falls. The prominence structure induced by the different response types, in a given discourse configuration, give us a baseline for the different uncertainty based scales we have been looking at. Looking at things in this way, we see that responses that only indirectly addressed the current QUD, were perceived as being non-final regardless of whether they exhibited a terminal rise or not. This indirectness also generally reduced the perceived strength the responder's opinion about the utterance it is responding to (reflected by the credibility and sureness ratings). This makes sense in pragmatic terms: if the responder was in the position to make an absolute affirmation or contradiction, they would have made a direct response. The methodological take-home point is that before attributing meaning to accent types, we need to take into account the baseline that is produced by the discourse configuration itself.

Previous studies have indicated that subjects sometimes perceive rises where there are none given the right prosodic and semantic context (Studdert-Kennedy and Hadding, 1973). So, it is not implausible that the lack of a rise was interpreted as a reduction in the prosodic gesture rather than an underlyingly different accent. Further perception studies testing whether subjects actually heard a rise or not would be needed to make this point. However, it seems reasonably clear at this point that

the salience of a terminal rise is conditional on discourse expectations. Another way to put this is that if the contribution of the terminal rise is parallel to what is already evoked by the discourse configuration, it's being there will not be very salient. If the response type expresses something orthogonal to what is expressed by the rise, e.g. a direct agreement declarative, then presence of a rise will need to be accommodated into the utterance interpretation.

Assuming that rises signal that the discourse has not come to a good stopping point, the prediction is that response types that also have this flavour will show more variation in the terminal rise characteristic in production than response types that don't. However, we would expect this variation to be marginalized out in perception on these epistemic/uncertainty type scales. We would expect to see bigger effects for response types with non-rise like meanings. These hypotheses fit what we have seen for the indirect responses as well as the cue word data. It is also the broad pattern for direct declarative statement responses versus declarative questions. However, a more fine grained look at the relationship between production and perception would be required to really determine the sensitivity of these move types to prosodic variation with respect to discourse attitudes.

This analysis implicitly assumes then that terminal rises be treated as separate to the contribution of the accent. Both tell us something about how an utterance slots into the current discourse. The next section looks at how these findings can be translated into semantic and pragmatic terms and the implications of doing so. To do this we look a bit more deeply at where the results from the current work lie with respect to the types of observations that have driven analyses of the fall-rise accent in the past.

6.3 The Proper Analysis of the Rise in Fall-Rise Accent

In this section, I attempt to reconcile the data collected in the current series of experiments with previous analyses of the fall-rise accent. Several of these have been already been brought up in the last two chapters with respect to the questions of whether the terminal rises, and nuclear accent shape more generally, can be interpreted as labelling units in IS category, dialogue act, or sentence type terms. The answer to these questions appears to be no. Rises instead tell us something about the discourse structure, i.e. they reveal expectations how those units fit in with other units. This obviously casts doubt on the robustness of previous analyses which attempt match intonational components with IS categories (Jackendoff, 1974, Gussenhoven, 1984, Steedman, 2000, Büring, 2003). The other approach is to analyze the fall-rise as

signalling some relationship between the proffered content of the fall-rise utterance and some other salient proposition in the discourse. Unsurprisingly, analyses of this relationship are centered on notions of uncertainty and unresolvedness (Ward and Hirschberg, 1985, Constant, 2007, Wagner, 2010).

Both IS category based and attitudinal analyses are at core about the treatment of contrast and alternatives, and by extension the relationship between the representation of these aspects of meaning and the perception of attitudes like uncertainty. As such there is significant lot of overlap between the two approaches. In the following, we look at the motivations behind these different analyses, their advantages and their flaws, in order to shed light on what semantic and pragmatic infrastructure is required for a more formal analysis of prosodic meaning.

6.3.1 Fall-Rise and IS categories

Jackendoff's B accent

The basic data that IS category based analyses such as Büring (2003) draw upon comes from Jackendoff's (1974) account of why the wide scope reading of negation is preferred when the utterance is pronounced with narrow focus and a low terminal rise, as in the following.

(6) *All* the boys didn't come...

In this work, Jackendoff refers to fall and fall-rise accents as A and B-accents.[1] This account attempts to draw a parallelism between the prosody of the previous sort of negative-quantifier utterance with pronunciations of the non-focus accent that are possible in responses to wh-questions, as in (7b) and (8b), which we can recall are the key data used in Büring (2003) (cf. Section 5.4.1).

(7) a. What about Fred? What did he eat?
b. *Fred*$_B$ ate the BEANS$_A$

(8) a. What about the beans? Who ate them?
b. FRED$_A$ ate the *beans*$_B$

In Jackendoff's theory of focus, a falling accent marks a dependent variable, while a fall-rise marks an independent variable. He analyzes the information structure in

[1] Jackendoff (1974) references 'Bolinger's B-accent', however it is clearly not what Bolinger (1989) describes as a B-accent. Instead, it appears to be an AC accent in Bolinger's terms.

terms of the new and the presupposed. The 'presupposition' is formalized as an open proposition representing a 'mapping from a set of values defined by the variable x, marked with a B accent, into a set of values defined by the variable y, marked with an A accent...' (pg 262). So, we can basically take this presupposition to be the IS ground, and the variables to represent contrastive elements, with an A accent mapping to the IS focus.

(9) *Gwendolyn*$_B$ analyzed the SPECTROGRAMS$_A$
   a. P(x) = {y: x analyzed y}
   b. P(Gwendolyn) = {spectrograms}

Example (6) is then analyzed as having only a B accent with the tail stretched out to the utterance end, and no explicitly marked independent (A marked) variable. In this case, he proposes using affirmation/negation as the dependent variable, i.e. polarity/verum focus. So, the open proposition maps the quantifier *all* to falsum. The overall effect is to remove negation from the presupposition/ground.

(10) *All* the boys didn't come...
   a. P(x) = whether x of the boys came
   b. P(All) = whether all the boys came = false
      $\rightsquigarrow$ not(All of the boys came)

This analysis presents the key ingredients for the IS based analyses that have followed. However, various other works have pointed out problems in this theory, especially with respect to the scope disambiguation by intonation facts that originally motivated Jackendoff's theory. Liberman and Sag (1974) note that, while this approach removes negation from the presupposition, it doesn't actually explain the preference for wide scope of negation. Moreover, they point out that negation can take narrow scope given the right context. Ward and Hirschberg (1985) note that a main motivation for this analysis is to exclude *not*, *only* and *even*, i.e. things that associate with focus, from appearing in the presupposition. However, in (11a) it seems that the response 'Bill doesn't like it' presupposes *x doesn't like San Francisco*.

(11) a. How can anyone with any sense not like San Francisco?
   b. *Bill* doesn't like it...

The problem here seems to arise from the use of the term presupposition to refer to backgrounded information. It clear that *somebody doesn't like SF* is not a 'normal' presupposition in the sense of being a common ground constraint or well-definedness condition (Stalnaker, 2002). Once we instead move from the notion of presupposition

to IS ground, we can see that what is really being represented here is the type of question being answered by the utterance. That is, question answer congruence.

From this QAC perspective, we get a better idea of the felicity conditions for the accents. For example, in (12), the A accent on *Fred* is felicitous with a negative wh-question, while the B-accent is not. However, the B-contour in other circumstances seems to implicate some sort of correction, as in (13). This suggests that if it does answer a question, the B-accent version answers a whether question. So, by virtue of being a contradiction and *Fred* being given, the IS focus in this case is the negative polarity.

(12) Q: Who doesn't write poetry in the garden?
A: FRED$_A$ doesn't write poetry in the garden.
A': ?*Fred*$_B$ doesn't write poetry in the garden...

(13) A: Fred writes poetry in the garden.
$\rightsquigarrow$ MAX-QUD =whether Fred writes poetry in the garden.
B: *Fred* doesn't write poetry in the garden...
(*Bill* does)

In fact, the *how*-question in (11a) is basically a biased/rhetorical question. So the B-accented response seems to really be a correction to an implicit whether-question.

(14) A: No-one doesn't like San Francisco $\rightsquigarrow$ Everyone likes SF.
B: *Bill* doesn't like San Francisco....
a. P(x) = whether x likes San Francisco
b. P(Bill) = false $\rightsquigarrow$ not(Bill likes San Francisco)

The interpretation of quantifier cases like (10) is then guided by the contradictory nature of such utterances. It makes sense that contradictions would induce *whether*-questions since contradictions apply to entire entire propositions. If we take the contradicted statement to assert universality the contradiction has to be interpreted with negation taking wide scope. That is, whether or not negation takes wide scope depends on the whether-question induced by the context. In the following we get the fall-rise contradiction contour of Liberman and Sag (1974) with high scope for *all*.

(15) A: John just told me all the rats escaped!
B: What? *All* the rats didn't escape...
B: Look, they're all here. He was just messing with you.

The crucial points of this sort of analysis are that (i) accent placement indicates the shape of the question under discussion, (ii) accents mark variable (alternative evoking) units, (iii) fall-rise accents don't mark IS focus. The production and perception data reported above fit with the first two points, but not the last. It seems we need some more structural notions to pin down what is happening in productions bearing this contour.

Büring (2003): Contrastive Topic

As discussed previously, one of the stylized differences between pair-list examples with nuclear fall-rise accent, e.g. (7b), and one that simply has two falling accents is that the former suggests that the preceding question wh-question has not been completely answered. This is the central point addressed in Büring's (2003) analysis of the fall-rise as being in a 1-1 relationship with contrastive topic category. This focus on incompleteness differentiates this analysis from that of Jackendoff (1974). The basic structure used to model this incompleteness in Büring's account comes from the theory of discourse/information structure set out in Roberts (1996), where discourses are modelled in terms of hierarchically related questions and answers. That is a *discourse tree* where nodes represent moves, with one representing the current question under discussion.

The key point for Büring's analysis is that a question node may dominate other question nodes as long as those subquestions are relevant in resolving dominating questions. Any subtree rooted in an interrogative is, thus, called a *strategy*. Büring argues that contrastive topics indicate/induce a strategy. So, the answers in both (16) and (17) indicate strategies with respect to the superquestion *Who ate what?*. The placement of the B and A accents determines, roughly speaking, what the strategy is about. The BA ordering in (16) signals that the speaker is answering the *Who ate what* superquestion on a eater by eater basis, while the AB ordering in (17) signals that the answers are 'sorted' by food. The main innovation of Büring's treatment from that of Roberts (1996) is the claim that contrastive topic invokes its own special semantic value, which together with the Rooth (1985) style focus semantics produces a set of questions rather than a set of propositions, e.g. (16c), (17c).

(16) a. What about Fred? What did he eat?
b. Fred_B ate the beans_A
c. $[\![\text{Fred}_B \text{ ate the beans}_A]\!]^{ct} = \{\text{Who ate x? } | x \in D_e\}$
$= \{\{y \text{ ate } x \mid x \in D_e\} | y \in D_e\}$

(17) a. What about Fred? What did he eat?
b. Fred_A ate the beans_B

c. $[\![\text{Fred}_A \text{ ate the beans}_B]\!]^{ct} = \{\text{What did y eat? } |y \in D_y\}$
$= \{\{y \text{ ate } x | y \in D_e\} | x \in D_e\}$

So, we can basically see this analysis as bringing together the notion of question answer congruence with Jackendoff's idea that the different accent types mark different types of variables: one to set the topic, another to set the focus.

Büring further argues that CT marking is *necessary* when a subquestion is implicit (although it is optional when the question is explicit). The argument for the necessity of CT/B-accent marking comes from examples like the following, where deaccenting 'female' seems infelicitous.

(18) What did the popstars wear? (Büring, 2003, ex.(27))
a. The female$_B$ popstars wore caftans$_A$
b. ?The female popstars wore caftans$_A$

This sort of data is dubious on two counts. First, the infelicity arising from not having a prominence on 'female' can easily be attributed to metrical constraints, rather than semantic ones.[2] For example, the sentence is quite long for there not to be any accents in the pre-focus region, and 'female' is not given which makes it a candidate for rhythmic prominence without the provoking contrast. Second, putting an A accent on 'female' is felicitous and still seems to give off some feeling of incompleteness by virtue of the fact that it is a partial answer. It's just doesn't sound as marked as the B accent version.

More generally, since Büring also posits a 1-1 relationship between CT and the fall-rise accent, this becomes a very strong claim about intonational meaning that is not supported by the empirical data from the current study (cf. also Calhoun (2007)). We

[2]To be honest, I disagree with many of Büring's claims of infelicity outside the ones that follow Jackendoff's original contrast. Büring uses the following example to argue that BA contours are infelicitous where the response is a complete answer to the question. However, the following seems perfectly fine to me, for example, a scenario where it was somewhat unlikely that Alessa would be doing the dumping.

(i) a. A: Did Knut break up with Alessa? Büring (2003, ex. 32)
b. B: No, Alessa$_B$ broke up with Knut$_A$

I would guess that people *would* find the AB version infelicitous, but I would suggest reflex of what the lack of QUD resolution associated with final rise and the fact that speaker B has actually resolved the question, which is something independent of whether there are strategies at play. A more general problem is the fact that, as noted by Roberts (1996), all questions in a discourse have a superquestion, e.g. *what is the world like?*, so shouldn't a strategy always be accommodatable? It seems to me that the answer is should be yes.

have seen that indirect responses, which much generate a strategy in this framework, are not always produced with a fall-rise, and similarly fall-rise accents can fall on IS focus. However, the idea that the implication of unfinished business associated with partial answers/indirect responses is a by-product of strategy marking seems to fit the data quite well once we take the burden off the shoulders of the fall-rise accent. For example, the following cases where there is another prominence the post-verum region seem to fit broadly into Büring's notion of subquestion: the utterances are interpreted as answering relevant *whether* questions. However, it doesn't seem that a fall-rise accent is necessary or even appropriate for the question/substrategy analysis.

(19) Did the girls do what they were supposed too?
   a. Well, Laura didn't write the report but Marianne did meet with Benny.
   b. {Whether Laura wrote the report, whether Marianne met with Benny,...}

In general, there seem to be many other intonational ways to get the same strategy evoking effect which don't involve a terminal rise. For example, placing a scoop accent (L*+H) on *female* in (18) seems to be another salient way to imply that other subquestions remain unresolved.

Büring takes care to point out that his notion of contrastive topic differs from the usual understanding of topic or theme. However, its not exactly clear what an operational definition of contrastive topic might be. From the point of view of Büring's framework, we would take contrastive topic to be the whatever a marks a strategy. Note that it seems completely felicitous to change a strategy mid response, e.g. (20).

(20) a. A: Who ate what?
   b. B: Well, Fred_A ate the beans_B, but Marianne_B ate the Jalapenos_A
   c. A: Oh really? I've always liked Marianne better than Fred.

Impressionistically, the B-accent on *Marianne* in (20) simply highlights some special salience, perhaps unexpectedness, with respect to the fact that Marianne ate jalapenos (e.g. perhaps Alice and Bert are trying to identify who out of their friends likes spicy food). It also seems possible to have BA type contour on verum focus examples, e.g. (21).

(21) A: Did Mary do her chores?
   B: Well, she didn't vacuum but she did_B do the LAUNDRY_A!
   A: Oh my god! Mary never does the laundry!
   a. {What did she do?}

	+agree	-agree
theme θ	L+H*	L*+H
rheme ρ	H*	L*

Speaker S	Falls: LL%, HL%
Hearer H	Rises: HH%, LH%

Table 6.1: Steedman (2007): Tone meanings for English

b. {Whether she did X}

It's not clear in this case how to apply Büring's calculation of question set. Since the A accent (focus) falls on the object, the sub-question set has an object *what*-question schema. However, we are then left with the problem of how to interpret the B-accent on *did*. Instead, it seems that the immediately dominating subquestion should be the *whether* question. Again, here the B-accent does seem to be signal some sort of salient contrast. However, the inferred discourse structure suggests that the B accented *did* should be interpreted as the IS focus.

So, it seems the strongest result we can take away is that a fall-rise accent is *one way* to signal the discourse configuration usually associated with contrastive topic. That is, one where the response doesn't directly resolve the current question under discussion.

Steedman (2000)-(2007): Theme and Rheme

While Büring's account deals with the fall-rise accent as a whole, the analysis suggests that the locus of meaning difference lies in the rise of the fall-rise, since this is what separates the B and A accents. This clearly doesn't work out with the data from our experiments. The framework of described in Steedman (2000) and Steedman (2007), on the other hand, puts the burden on pitch accent shape modulo the rise characteristic. So, we'd like to see if this would work any better. Steedman provides a compositional theory mapping intonational units identified in Pierrehumbert and Hirschberg (1990) to information structural categories. The basic information structural categories in this theory are theme and rheme, which broadly correspond to the notions of IS ground and focus respectively.

Like Büring (2003), this approach makes strong claims about the mapping between intonational form and logical form, summarized in Table 6.1. In this model, L*+H and L+H* both mark informational units as part of the theme, while L* and H* mark rhemes. High accents mark information that is agreed upon, i.e. legally in the

common ground, while low accents mark contentiousness with respect to this.[3] The (dis)agreement can be directed towards either the theme or rheme and informational units. Boundary tones attribute informational units to either the speaker's or the hearer's knowledge from the point of view of the speaker.

In this system, the fall-rise accent (L+H*LH%) marks agreed upon themes that the speaker takes the hearer to be responsible for. Similar to Jackendoff (1974), the claim here is that utterances with only this contour have no IS focus. Instead they are used to mark a change of theme. The explanation for the implication of unfinished business associated with these sorts of utterances is that the speaker accepts the theme but does not take responsibility for the rheme, which in turn implicates a lack of confidence. However, we can't use this sort of reasoning when there is an explicit rheme, as is the case for the indirect statement responses we have been analyzing. Steedman notes that themes are not usually accented unless they stand in contrast with something else. So, while Steedman does not directly address the implication of open-endedness associated with AB/BA marked utterances, it seems that he would account for it via the fact that accenting a theme invokes a contrast set from which similar strategy type inferences could be made. However, the interpretation of H% tone then becomes problematic. We should see a difference in the certainty/sureness ratings between indirect responses with different terminal rise characteristics in the perception experiment since the former indicates that the hearer is responsible. We can remove this problem by assigning H% the non-finality meaning instead. This brings the analysis of the indirect response data very similar to that of Büring (2003) having discounted the idea that the boundary tone is the distinguishing factor.

The idea that IS unit type is distinguished by pitch accent type opens up another dimension of comparison with the low target pitch accents which is problematic. As noted above, low accents are taken to indicate that the associated informational unit is contentious. That is, unlike the B-accent, Steedman takes L*+HLH% to indicate that the theme is not agreed upon. However, the association between L* and disagreement seems a bit tenuous. We can note that this encoding of contrariness in L*+H is at

[3]There is a bit of a change from Steedman (2000) to Steedman (2007) with respect what role the common ground plays in this, i.e. the latter talks about agreement rather than CG. This makes sense since the disagreements over theme and rheme may be more about the state of other discourse structures like the QUD. Also, the glosses get quite weird otherwise. For example, the idea that the speaker does not wish to put their lack of knowledge in the common ground in the following similar example seems quite far-fetched.

(i) a. H: Is it raining? Calhoun (2007)
b. S: I don't KNOW
L* LL%
c. "I do not make it common ground that I don't if its raining"

odds with the fact that fall-rise contours, also characterized by L*+H LH% by Ward and Hirschberg (1985), can be used in affirmative scenarios. In (22), for example, it seems that we can get L*+H even on the verum focus element in an affirmative scenario.

(22) Alice presents a cake heavy menu to Bert

a. A: I think Mary is going to be very happy with the menu.

b. B: She does love to eat cake
L*+H L*+H LH%

If such strong mappings from intonation to meaning are real we would expect this to be clearly reflected in speech. We would expect that we could categorize FR contours depending on whether they were in agreement or disagreement based on whether the accented syllables were really scooped: L*+H type accents would be used in contradictory contexts, while L+H* accents should be use in affirmative ones. However, we did not see this sort of accent distinction in the production data comparing indirect contradictions and agreements. However, even if we loosen up the accent definitions, it seems to get this we still need some extra structural representation like that provided by Büring's strategy analysis.

Prominence placement, not Pitch Accent Shape

Both Büring's and Steedman's accounts suffer an overly tight binding between an abstract notion pitch accent/boundary tone types and meaning.[4] It seems safe enough to assume that contrastive topic and focus, in a broad sense, are constantly required in real conversational speech. So, it seems unlikely that the former would be so tightly bound to such a 'flamboyant'[5] intonational event as a fall-rise accent. In this vein, Steedman notes that the mapping to those specific tones is not crucial for his theory as a whole, since the goal of his work is really to establish the place of IS in the grammar rather than to provide a fully robust prosodic theory. In fact, he suggests that many other contextual factors are involved in the IS labelling task.

With this in mind, it seems that the interpretations that have been associated with the fall-rise accent/CT arise from discourse/IS structural conditions. This does not appear to be prosodically signalled by pitch accent shape, but instead by accent placement interpreted with respect to the expectations of question answer congruence.

[4]Steedman (2000): 'It is important to be clear from the outset that there is no single definitive characterization of the components of intonational contour, much less a definitive theory of their information structural meanings.'

[5]Steedman's word!

This is the sort of approach argued for by Calhoun (2007), who argues that IS units map to prominences in a probabilistic way. That is, while IS foci usually align with the strongest element in the metrical structure, the mapping is conditional on other structural factors such as syntactic and phonological features. Deviations from default structure are usually interpreted as marks of contrast and in general contrastiveness is marked by increased prominence.

In support of her theory, Calhoun provides empirical evidence against the idea that themes and rhemes are distinguished by accent type via production and perception studies. Instead, rhemes are found to be relatively more prominent then themes (the only consistent distinction was in terms of pitch height). Height differences are also affected by relative position of accents with respect to the nuclear prominence (we see reduction in the post-nuclear region). The current work differs from Calhoun's in that we vary the discourse context primarily in terms of question answer congruence, while she looks more at contrast across sentences (cf. Section 5.4.1). Nevertheless, like the current work, she does not find perceptual evidence that rises are associated with contrastive themes/topics.

The sort of discourse configuration that we appear to be dealing with in cases where fall-rise accents are elicited are those where the response does not directly resolve the question under discussion. The production and perception facts suggest that we want to dissociate the contribution of the rise from this. However, the contribution of the rise in those circumstances is parallel to what would be conveyed by the utterance without the rise and so it is not salient. The key ingredients gleaned from Jackendoff (1974) can be revised as follows:

(23) a. Accent placement indicates the form of the question under discussion.
b. When alternatives projected by a response, via accent placement, don't match the alternatives projected by the question under discussion, the speaker will infer that a strategy is being played.
c. Rises signal discourse non-finality.

6.3.2 Intonation and Attitude: The Other Proposition

So, we've seen how an analysis of the fall-rise accent goes from an information packaging point of view. How does this compare to analyses which explain the interpretation of the fall-rise utterances as commenting on propositional level content? These accounts assume that the contour provides an extra layer of meaning which interacts with the semantic content of the utterance, and as such present more of a semantic operator approach. The implementations vary from the association of the contour with a specific attitude (Ward and Hirschberg, 1985), to attempts to analyze this

this contour as a focus sensitive quantifier (Constant, 2007). However, like the IS approaches reviewed above, these accounts all start from the intuition that the fall-rise says something about some proposition other than the one actually uttered. However, the treatment of alternatives is quite different. In this section, I will attempt to argue that the information packaging point of view provides a more sound basis to look at the interpretation of sentential prosody.

Ward and Hirschberg (1985)-(1992): Scalar Uncertainty

Ward and Hirschberg (1985, W&H) present an account of the rise-fall-rise contour, as conventionally implicating uncertainty with respect to values on a scale or the scale themselves. According to Ward and Hirschberg (1985), (R)FR is characterized by one or more scooped accents followed by a terminal rise, i.e. L*+H LH%.

(24) a. Help! I need the name of a disease that isn't incurable that starts with a sonorant!
b. *MelaNOma* isn't incurable...

W&H claim that scooped syllables mark constituents as variables in an open proposition and that the values of these variables can be seen as lying on some sort of scale. They suggest that speakers may be uncertain about evoked scales in three ways:

(25) Ward and Hirschberg (1985, (51))
a. I. Uncertainty about whether it is appropriate to evoke a scale S at all.
b. II. Uncertainty about which scale S to choose, given that some scale is appropriate.
c. III. Given some S, uncertainty about the choice of some value b on S.

Ward and Hirschberg (1988) revise their generalization for FR to capture the fact that it can also express incredulity. These examples are problematic because, unlike the examples above, they don't really seem to have the force of an assertion, but appear more like questions. Hirschberg and Ward (1992) suggest that in these cases the speaker is expressing a sort of disapproval about a value/scale that has been proffered by someone else.[6] Thus, the reasoning is that discourse context together with a more 'involved' prosody (e.g. greater pitch range and maximum amplitude) allows conversational participants to distinguish these interpretations.

[6] 'It is the case that the speaker thinks the scale/value is not appropriate.'

(26) a. I'd like you to be here at 11am (Ward and Hirschberg, 1988, (2))
b. *Eleven* in the morning...!

So, the core of this analysis is that when a speaker uses FR, they are making some comment about a value with respect to a scale.[7] The way in which W&H suggest to determine the relevant scale relies on the identification of a salient proposition P in the context that is not the FR utterance.

(27) Ward and Hirschberg (1985, (31))
a. Some proposition P becomes salient in the discourse.
b. A value b_1, referenced by the utterance of P is perceived as salient (P then can be represented as $P(b_1)$).
c. b_1 is perceived as lying on some scale S.
d. $P(x)$ then represents the open proposition formed by substituting a variable x for b_1 in $P(b_1)$.
e. A speaker instantiates $P(x)$ with some b_2 that co-occurs with b_1, on S (and may be identical with b_1).
f. The resulting $P(b_2)$ is affirmed or denied by the speaker with FR intonation

This analysis does not put any restriction on the relationship between b_1 and b_2 except that they are on some scale. The difference between the open proposition P referenced here and that used in Jackendoff (1974) is that P is not the utterance carrying the FR accent, but some other salient proposition. In the following example suggests that we can take the salient proposition P to derive from MAX-QUD, or some other dominating question under discussion. Here we picking MAX-QUD gives us P = *Mary robbed the candy store*. This gives us an open proposition which is a template for the properties of Mary.

(28) a. A: Do you think that Mary robbed the candy store?

[7] Pierrehumbert and Hirschberg (1990) go on to claim that this evocation of scalarity comes from the bitonalness of (L*+H) rather than the contour as whole. The only example they give for L% version is a vocative:

(i) Leo (Pierrehumbert and Hirschberg, 1990, (29))
L*+H HL%

However, it seems that you can just as easily invoke a scale with a H* type accent, so the scooped nature of the rise doesn't actually seem to be that crucial. So, it seems we are just dealing with FR rather than RFR. Note, however, that the scooped utterances would seem to contradict P&H's claim that L* accents don't predicate.

b. B: She DOES like *chocolate...*
   (i) P = Mary x-ed
   (ii) b_1 = robbed a candy store, b_2 = likes chocolate
   $\rightsquigarrow$ Mary likes chocolate $\rightsquigarrow$ Mary has a motive

It appears that the connection between the question and the FR-response in (28) is that liking chocolate is a possible motive for robbing a candy store. Hence liking chocolate is a piece of evidence which supports an affirmative answer to MAX-QUD. The relevant scale then seems to be how well properties of Mary indicate a robbery motive. The uncertainty seems to be about whether this piece of evidence is good enough to put a guilty label on Mary (W&H's type III uncertainty). However, if the fall-rise is the thing signalling the uncertainty, we would then assume that having a non-rising accent in the same position would not signal uncertainty. For our perception experiment above, we would expect to see a difference in the sureness ratings based on the final rise characteristic.[8] So it seems the feeling of uncertainty is coming from somewhere else.

Similarly, although we can induce an open proposition P from A's statement in the following, we still need IS notions to explain the accenting facts, and the presence of the verum 'do'.

(29) a. A: We haven't got anything connecting Marianne to Lenny
   b. B: Marianne did meet with Lenny...

(30) a. P = X connects Marianne to Lenny
   $\approx$ What connects Marianne to Lenny?
   b. b_1 = Nothing, b_2 = Marianne met with Lenny

The necessity of a scalar structure in these situations is also doubtful. We see in the following that with a little context, the infelicitous example (31) becomes fine, when there is enough information for the final rises to make sense.

(31) a. A: Did they have a boy or a girl? Ward and Hirschberg (1985, (17))
   b. B: # They had a *boy...*

(32) a. A: I'm not sure what colour onesie to buy Penny and Desmond.
   b. A: Did they have a boy or a girl?
   c. B: They had a *boy...*
   d. B: But Penny's militantly against gender stereotypes, so I wouldn't buy blue.

[8] Recall the question was 'How **sure** does B sound about whether A's statement is true or false?'

(33) Q = What properties of Penny and Desmond are relevant to selecting a onesie.

In this case, we can interpret the rise in the fall-rise as indicating that A needs to consider more subquestions than just the gender of Penny and Desmond's baby when addressing the superquestion of what colour onesie to buy.

Once we put these discourse structural and IS notions in play, the analysis looks a lot like the IS based one. Like Jackendoff (1974) this analysis associates the FR accented constituent unit with a variable in an open proposition. This gives us the template for a potentially implicit question under discussion. The difference is that W&H's analysis doesn't assume that this question should be derived from the utterance carrying the FR accent. However, this basically still centers around the notion of question answer congruence. So, in a way we can see the IS/strategy analysis is just a natural clarification of the approach suggested by W&H's, given sharper discourse notions.

Constant (2007): Focus Semantic Value

Constant (2007) presents a different way to connect W&H's analysis of FR to the notion of focus. In this analysis FR[9] is treated as a focus sensitive quantifier that acts on the level of Conventional Implicature (CI, Potts, 2005). In this account FR parallels *only* in that it acts on focus alternatives generated in the sense of Rooth (1985). The difference between *only* and FR is then that *only* asserts that all focus alternatives are false while FR utterances leave some unresolved. The unwillingness to resolve such alternatives is often attributed to a lack of knowledge, however it may be for other reasons, e.g. the speaker simply doesn't want to reveal what he knows. This is illustrated in (34).

(34) A: Did your friends like the movie? (Constant, 2007, (23))
   a. B: *John* liked it... I dont know about the rest of them.
   b. B: Only JOHN liked it. # I dont know about the rest of them.

(35) a. $[\![only(\phi)]\!] = \forall p[(p \in [\![\phi]\!]^f \text{ and } \hat{}p) \rightarrow (p = [\![\phi]\!]^o)$
   b. $[\![FR(\phi)]\!] = \exists p[(p \in [\![\phi]\!]^f \text{ and } ([\![\phi]\!]^o \not\rightarrow p)$

Constant argues that neither *only* nor FR can quantify over clauses that are *alternative dispelling*, defined in (36).

[9] Constant calls this RFR because he thinks that there is some import with respect to there being an rise in the contour. I will stick with FR since it does seem more commonly used in the literature, though he may have a point.

(36) The focus of clause ϕ is alternative dispelling iff the proposition denoted by ϕ resolves all alternative propositions generated by the focus. Using Rooth's notation: The focus of ϕ is alternative-dispelling iff $\forall p \in [\![\phi]\!]^f : [\![\phi]\!]^o$ resolves p.

(37) A proposition p resolves a proposition q iff either $(p \rightarrow q)$ or $(p \rightarrow \neg q)$

So, in this account, use of FR is infelicitous when the proffered content of the utterance settles all alternatives. Certain lexical items are by their nature alternative dispelling. For example, universal quantifiers in non downward-entailing environments or adjectives that denote end points of scales, e.g. *perfect*. For Constant, this leads to vacuous quantification.

(38) a. ??*All* the boys came...
b. ??Only ALL of the boys came

(39) a. #The food was *perfect*... (Constant, 2007, ex. 26)
b. #The food was only perfect.

In general, we can see this as a refinement of Ward and Hirschberg (1985). The other salient proposition, P, and scalar values, b_1 and b_2, are determined by the focus semantic value of the utterance. Along the lines of Ward and Hirschberg (1985), Constant argues that FR provides a conventional implicature in the sense of Potts (2005). The main arguments for this come from the non-defeasibility of the meaning associated with FR and its non-embeddability/speaker orientation. The scalar component of W&H's analysis is then subsumed by a scalar interpretation of focus alternatives, and the feeling of uncertainty arises as a conversational implicature.

In Constant's analysis the fall-rise is treated as one meaning unit and given a lexical entry as such. However, unlike IS labelling approaches like Büring (2003), the FR meaning can parallel the meaning of the equivalent non-rising utterance. For example, a non-rising utterance could just as easily have unresolved alternatives.

(40) Do you think Jack is home?
a. The *lights* are on...
b. The LIGHTS are on.

So, the contribution of the fall-rise in this sort of indirect response situation is neutralized by the discourse configuration. This means that FR is not necessary, but it is sufficient for signalling that alternatives are unresolved.

Like W&H's analysis, this account does not say anything about the relationship between FR and the QUD, IS topic/focus, or prominence placement. The assumption being that primary determiner of the distributional facts is the focus semantic representation. So, it suffers from the same drawbacks as that analysis in terms of explaining how this accent relates to other prosodic features in the utterance. Unlike W&H's analysis, however, Constant makes a stronger claim in analyzing it as a focus sensitive quantifier. However, we will see that there are some limitations in how far we can go only looking at the focus semantic value without referencing discourse and information packaging notions.

Focus Alternatives vs Strategies

The focus semantic approach leaves the determination of alternatives up to other contextual factors. This means that whether or not a proposition is interpreted as alternative dispelling still depends on discourse structural considerations. As we saw previously, having an explicit superquestion can change whether FR is felicitous, since this can change the availability of different alternatives. For example, *perfect* is licit under FR when we consider properties of *the food* relevant to evaluating whether the speakers would go back to a particular restaurant again.

(41) The food was *perfect*...but it wasn't really worth the money.
   a. Alts = {The food was perfect, The food was worth the money, The food was cheap,...}

In these sorts of cases we can see that the alternatives suggest some sort of strategy for evaluating some higher level question under discussion. This is more salient in our verum examples. Here we see the focus alternatives mirror *whether*-strategies.

(42) A: Do you think Mary robbed the candy store?
   B: She does come from a *crime* family...
   a. $[\![p]\!]^o$ = She comes from a crime family
   b. $[\![p]\!]^f$ = {She comes from a crime family, She likes candy, she has a history of larceny,...}

Assuming that such VP alternatives are available from the discourse, it might seem that the focus alternatives and a QUD/strategy approaches are more or less the same, i.e. that focus alternatives implicitly hold a strategy. So, one argument in favour of framing the theory in terms of focus alternatives is that it provides an elegant parallelism between FR and the analysis of focus sensitive operators like *only*.

However, the parallelism between *only* and FR is not perfect. Constant argues that both FR and *only* quantify over propositional alternatives. So the focus semantic value of an FR or *only* utterance must not be alternative dispelling in order to avoid vacuous quantification. So in the following the claim is that both FR and *only* are illicit because of the alternative dispelling quality of *perfect*.

(43) How was the food?
   a. ?The food was *perfect*...
   b. ?The food was only *perfect*...
   c. Alt. = {The food was ok, the food was good,...the food was perfect}

However, adding an additional prominence on *food* rescues the FR utterance, while the *only* version still seems odd, i.e. (43). At first glance this is because the ordinary semantic value now doesn't resolve the alternatives as is required by *only*. However, when the food is evaluated to 'only good', the utterance is fine.

(44) How was the food?
   a. The *food* was *perfect*... (The wine was not good)
   b. ??The *food* was only *perfect*.
   c. The *food* was only *good*.
   d. $[\![$The *food* was *perfect*$]\!]^f$: {The food was okay, the food was good, the food was perfect,... The wine was ok, the wine was good, the wine was perfect...}

One could say *food* is the topic so the things it contrasts with do not go into the focus alternative set. It is not clear then why the food alternatives are allowed (in fact necessary) for the FR utterance, but disallowed for the *only* version.[10] Similarly, in a scenario like (45), it is very hard to use *only* to express that only one possible motive can be attributed to Mary, unless there is an explicit restriction. That is, this reading is possible when there's an explicit reference to strategy being used to evaluate A's question. This differs from the FR use, e.g. (42), where we had to assume that the motive (VP) type alternatives were salient. So, again, it does not appear that the alternatives picked out by FR are those that are picked out by *only* unless some coercion is involved. That is, the strategy component seems tied to the meaning of FR but not to that of *only*.

(45) A: Do you think Mary robbed the candy store?

[10]Constant uses this conception of double foci as generating exactly this type of multivariable alternative set in his later analysis of the interaction of FR with quantifier scope.

a. ??B: She only likes chocolate.
b. ??B: She only loves thieving.
c. B': Out of the possible motives, she only likes chocolate.

Besides this, a problem also arises with respect to whether FR can combine with *only*. In Constant's account FR should be evaluated after *only*. Since *only* resolves the alternatives, he predicts infelicity for (46).

(46) Only *John* liked it...

However, feelings of infelicity in the isolated case are resolved when we provide a dominant unresolved question which makes the *only* utterance part of a strategy, e.g. (47).

(47) A: Do you think the movie will be a failure?
B: Well, only *John* liked it... (and he's way outside the demographic)
(i) resolved: {whether X liked the movie}
(ii) open: whether the movie will be a failure. $\rightsquigarrow$ probably

So, this suggests that the unresolved alternative in Constant's analysis need not come from the focus semantic value of the FR-utterance, or at least the focus semantic value that would be associated with *only*.

Looking at the data from a strategy point of view also shed's light on cases where FR is used in situations that would usually count as alternative dispelling, but instead seems to insinuate that there are unresolved alternatives. That is, cases where FR is accommodated. For example, in (48) Bert basically asks Alice to resolve for herself whether Mary being miserable resolves Mary being sad.

(48) A: Is Mary sad?
B: She's *miserable*...
a. $[\![\text{She's miserable}]\!]^o$ = Mary is miserable.
b. $[\![\text{She's miserable}]\!]^f$ = {Mary is happy, Mary is sad, ..., Mary is miserable}

Similarly, FR on *d'uh* suggests that the existing evidence is sufficient to determine whether to accept the proposition *Marianne is in love with Lenny*. In this case, we treat *d'uh* like a cue word in that we assume that it comments on a proposition already in the discourse. We can analyze d'uh(p) as saying that p is evident.

(49) A: Marianne *is* in love with Lenny
B: *D'uh*...

a. $[\![\text{D'uh}]\!]^o$ = d'uh = That Marianne is in love with Lenny is evident
b. $[\![\text{D'uh}]\!]^f$ = {That Marianne is in love with Lenny is evident, That Marianne is in love with Lenny is plausible,...}

So, it seems that FR can be accommodated when an appropriate set of alternatives can be identified, even when the FR utterance itself is alternative dispelling. That is, when the immediate question under discussion can be identified as being part of a strategy. In the rhetorical cases, this strategy is invoked to suggest that the hearer figure things out for themselves. This may be a rather conventionalized use, but it still seem to draw from the same basic discourse components.

Wagner (2010): The Other Speech Act

The discussion above suggests that to understand what is going on with FR utterances, we need to more than just the focus semantic value of the utterance. However, it is still instructive to look at how far we can get in an analysis where the FR definition is further divorced from both IS packaging and F-marking/kontrast. Such an analysis is presented in Wagner (2010). The argument that the unresolved proposition does not need to be a focus alternative of the FR carrier comes from the observation that FR is the canonical contour for the response *d'uh*.

(50) A: OMG! Marianne is in love with Lenny
B: *D'uh...*

I have already suggested above that *D'uh* does actually provoke well formed alternatives. However, it is still useful to consider the implications of the much less restricted definition of FR Wagner gives. The assumption here is that the basic contribution of FR is a presupposition that the speaker might have said some other salient thing but they didn't. In the following let S and S' be assertive speech acts, we then have:

(51) $[\![\text{FR}_{\text{wagner}}]\!] = \lambda S \in D_S : \exists S' \in [\![S]\!]^g_a, S \not\rightarrow S'$ and asserting S' might be justified. S

Overall, we can see Wagner's account as something of a return to a whole contour type of analysis attributed to highly conventionalized intonational forms like the Contradiction Contour (Liberman and Sag, 1974). Given that this analysis basically abstracts over both the F-marking/kontrast and IS packaging components of prosodic meaning, it is unsurprising that Wagner's FR definition is very close to the notion of non-finality we have been attributing to terminal rises. The perception data also tells us that this feeling of non-finality is more or less inherent to indirect responses.

So, it seems that the connection between indirectness and FR is basically what is captured in Wagner's definition. To maintain relevance, the salient alternative has something to do with the current question under discussion (or some dominating question). In fact, we would expect the salient alternative to to be the canonical, i.e. direct, response projected by the QUD. However, selecting S' this way is problematic for the constraint that $S \not\rightarrow S'$ in examples like the following.

(52) A: Nobody brought a dessert!
B: Emily did bring a *meringue...*
(i) S = Emily brought a meringue
(ii) S' = Somebody brought a dessert.
$S \rightarrow S'$
(iii) S' = You're blind.
$S \not\rightarrow S'$

Here, it seems quite clear that S does entail S'. Choosing a contradiction variant like 'You're blind' might get around this. However, given the way FR is accommodated in other circumstances, it seems that for FR to have the proper effect the salient proposition needs to be accessible by the addressee. However, while a claim of blindness might be salient in B's mind it is doesn't seem salient for A (who we assume believes at that point that there were no desserts). Moreover, it seems that we still need more machinery to explain the distribution of FR. In the following an FR response is odd if we switch S and S' from above, even though in this case we do have $S \not\rightarrow S'$.

(53) A: Nobody brought a dessert!
B:??Someone did bring a *dessert...*
(i) S = Somebody brought a dessert.
(ii) S' = Emily brought a meringue
$S \not\rightarrow S'$
(iii) S' = You're blind.

In general, it seems that what licenses FR is not just the salience of another speech act, but the fact that B didn't use the canonical response to address the previous claim. So, we still need to bring in the discourse structural and IS notions to make some progress on what is actually going on prosodically.

6.3.3 Metalinguistic Uses

FR utterances have also been associated with incredulity. Examples like the following are problematic for analyses of FR that invoke unresolved or unspoken alternatives

propositions since they don't really seem to reference alternatives in the focus semantic value of the utterance.

(54) B: How did they get in?
A: There were marks that suggest that the windows were opened.
B: What?? The *windows* were opened...! I don't believe it!
Foc. Alt. = {The windows were opened, The doors were opened, Someone burrowed through the basement,...}

The example above appears to lack the force of an assertion. This leads Constant to analyze them as metalinguistic in nature, excluding them from his analysis. However, FR also seems to be licit in incredulity scenarios were the FR utterance does seem to have the force of an assertion.

(55) B: I'm working on meningitis. How about you?
A: I'm working on melanoma. It's incurable, you know.
B: What? *Melanoma* isn't incurable...!
(i) FSV = {Melanoma isn't incurable, Menigitis isn't incurable,...}
(ii) Strategy ={Melanoma isn't incurable, Melanoma isn't incurable}
= whether Melanoma is incurable.

In (55) where Alice and Bert are talking about diseases they are researching. Since prominence falls on 'melanoma' we can assume this evokes alternatives in terms of other diseases. Now, while it's true that B's asserting melanoma's curability does not resolve whether meningitis is also curable, B's response doesn't really seem to say anything about alternative diseases one way or another. So it seems the focus semantic value doesn't help us much in this case.

However, a strategy does seem able to do the job once we note that what B's response does is pick out *melanoma* as the object for correction and ask the hearer to acknowledge that correction. We can think of this as a producing a polar (degenerate) strategy, where the prominence picks out the item to be corrected. The key difference between this example and (54) is the FR speaker acts as a source for the FR proffered content in the former but not in the latter.[11] This highlights the necessity of taking

[11] We can find cases that lack the force of assertion, but do seem to comment on alternatives, as in (i).

(i) a. B: The big the news is that Kate and James got married.
b. A: What?? *Kate* married James...
c. A: No! Surely you meant Juliet!
d. $[\![\text{Kate married James}]\!]^f$ = {Kate married James, Juliet married James,...}

Level	Clark (1996)	Allwood et al. (1992)	Clarification example
1	Attention	Contact	Did you say something?
2	Identification	Perception	What word did you say?
3	Understanding	Understanding	What did you mean by x?
4	Consideration	Attitudinal reaction	Should I put this into the common ground?

Table 6.2: Levels of Grounding

into account what type of dialogue move and hence what type of strategy is being played.

The question remains whether these examples should really be kept out of an analysis of FR. I would argue that the fact that metalinguistic negation is associated with this tune is an offshoot of what the core meaning of FR is rather than a separate entity. Stepping back a bit, we can see that metalinguistic uses of FR pop up when a *grounding* problem arises (Clark, 1996). Grounding problems are far from rare (Purver, 2004, Schlangen, 2004, Benotti, 2009, a.o.) and corrections happen at levels varying from the reception/acceptance of the acoustic signal to the propositional level (see Table 6.2). A check on grounding at the propositional level can result in the expression of incredulity or surprise.[12] In (56) B signals that John loving the movie is surprising by using FR to suggest that the *whether* question is still open, until he implicitly accepts the positive answer with the positive follow up.

(56) A: John loved that movie
B: He *loved* it... That's great! We're going to be rich!!
(i) open: whether John loved the movie

Inability to ground at lower levels produces the more familiar metalinguistic examples, as in (57).

(57) I don't like *[ae]pricots*... I like [ei]pricots
a. closed: whether I like [aeprikots]
b. open: whether I like apricots

While grounding related moves are by their nature discourse oriented, it is not clear that we want to put everything associated with such moves in the metalinguistic wastebasket. In general, it seems that a theory of prosodic meaning that *can* apply down at these lower levels is preferable to one that cannot.

[12]Like cue word *really*!

6.3.4 Strategies, Rises and Coherence

We identified two major types of analyses for the FR contour. The IS based approaches following Jackendoff (1974) have claimed that different pitch accent shapes select different IS category categories. From this perspective, the fall-rise has most saliently been connected to the notion of contrastive topic. Another strand of analyses, following Ward and Hirschberg (1985) take FR as commenting on a proposition with respect to some other salient proposition, rather than a sub-propositional category. A large part of the difference between these two approaches is the type of FR utterance these analyses are derived from. IS type analyses like Büring (2003) mainly deal with utterances in which one prominence is the IS focus and the other is in the ground, e.g. pair-list answers to wh-questions. The alternative proposition approaches like Constant (2007) mainly consider utterances with only a fall-rise accent. As such, we can see why there would be less of a focus on IS categorization of accents within an utterance. The indirect responses elicited in our production experiment, however, fall somewhere between the two.

The stylized fact that connects these different utterance types is that FR nuclear accents come with a feeling of incompleteness. The key point from the IS accounts that helps us understand why we get FR accents when we do and why we get the impression of incompleteness is the notion of question answer congruence. Once we look at the relationship between an FR utterance and QAC, we see that FR is licit when the utterance only indirectly addresses the current question under discussion. On the flipside, FR can be accommodated when there is a dominant question for which the current QUD can be interpreted as partially addressing. That is, when the speaker appears to be using a strategy, i.e. a series of subquestions, to address this live question. However, the production and perception data collected in the current work argue against associating strategies exclusively with fall-rise accents. Instead, all that seems necessary is that the response be unexpected given the form of the current question under discussion, and that the strategy induced from the IS structure of the utterance maintain discourse coherence.

A conventionalized type of strategy that works in many cases is a whether-strategy, where the speaker appears to weigh up various propositions in order to evaluate the QUD. This is the type of scenario where verum focus appears naturally.

(58) a. Is John a good person?
b. Well, he did rob that *bank*...
c. But then he gave the money to the poor.

For indirect responses to polar questions or statements a strategy interpretation is more or less inevitable, again for reasons of discourse coherence. Responses with final

falls are perfectly capable of signalling this discourse configuration. However, even though the rise seems negotiable, we see that placement of prominence has to fit with a plausible strategy or we do get a feeling of incoherence as in (60b).

(59) A: Do you think John's home? (Expected: {yes, no})
   a. B: The LIGHTS are on.
   b. B: The *lights* are on...

(60) A: Nobody brought a dessert.
   a. B: Emily DID bring a MERINGUE
   b. B: ?? Emily DID bring a meringue.

Pair-list answers to wh-questions have a better fit to the QUD than the responses to the statements/polar questions above. In the following, the FR pair-list utterance has the direct response form. In this case, it does seem that a rise is necessary to overcome the default interpretation of resolution. The case is the same for direct responses to polar question. Similarly, it seems that we need a rise to overcome deaccenting constraints from givenness.

(61) A: Who ate what?
   a. B: FRED ate the BEANS. (direct/expected)
   b. B: FRED ate the *beans*...

(62) A: Who ate the beans?
   a. B: FRED ate the beans. (direct/expected)
   b. B: ? FRED ate the BEANS
   c. B: FRED ate the *beans* ...

(63) A: Is John a good person?
   B: Yes, he *is*...

So, it would seem that the final rise facilitates a strategy interpretation. How necessary the presence of a rise is for this depends on how far the response is from being direct. This use of a final rise exactly fits the mandate of non-finality we have attributed to it. The rise in internal fall-rise accents can then be thought of simply as indicating the unfinished status of the utterance. What actually seems to be triggering the strategy interpretation suggested in these cases is the contrast suggested by the utterance internal boundary: putting the topic into its own phrase is unexpected.

(64) A: Who did Fred eat?

B: *Fred...* | ate the BEANS.

Taking IS categories into account allows us to say more about how the utterance fits in with the previous discourse. This is crucially what the alternative proposition type accounts can't tell us. So while those analyses based on the focus semantic value of an utterance can tells us something about the alternatives, we still need IS packaging notions to really figure out what the prosody is contributing.

6.4 Conclusion

This chapter investigated the contribution of terminal rises in fall-rise accents in terms of utterance interpretation. The production experiment described in the previous chapter found fall-rise accents to be consistently elicited in the IS ground of indirect statement responses. However, closer inspection of the data showed that not all accents in this position exhibited an actual rise. The perception experiment presented in this chapter investigated whether the presence or absence of a rise made a difference to how these responses were interpreted in terms of four attitudinal scales: credibility, sureness, non-finality, and engagement. The first three ratings reflect different types of uncertainty in the discourse that have been associated with fall-rise accents and rises more generally.

The results indicated that having a rise didn't really add to the perception of uncertainty beyond what we get from the discourse configuration, i.e. the uncertainty attached to being indirect response. We saw some indication of a positive trend with respect to tilt and convex for the non-finality scale, but the estimated effects of these features was not significantly different from zero. However, we did see significant effects of peak height for the engagement scale, indicating that the participants were actually attending to the prosody on some level. This presents more evidence against a strong characterization of IS units based on the final fall/rise characteristic of a pitch accent. From a methodological point of view, we see that in order to gain a robust understanding of the contribution of prosody on a particular meaning dimension, we need to take into account the baseline induced by the discourse configuration itself. We saw that indirect responses generally garnered less extreme values on credibility and sureness scales, but higher values on the non-finality scale. The latter fact indicates that indirect response configurations have some inherent non-finality. So, much like the cue word *really*, it seems that the contribution of the final rise is dominated by parallel cues from the discourse configuration in this case.

In this way the data from the production and perception experiments shed light on why fall-rise tunes have been associated with contrastive topics in the sense of Büring

(2003). Indirect responses, by virtue of not performing the canonical follow-up move, are exactly the type of responses that evoke strategies, where we see strategies as a way to maintain discourse coherence. We view the set of alternatives involved in strategies as a set of questions. This level of structure is what distinguishes this approach from analyses based on the Rooth style focus semantic value of an utterance, i.e. alternative propositions. Sometimes the form of the FSV and the relevant strategy project the same thing. However, examining the parallelism between FR and *only* bearing utterances we saw that this is not always the case. We see that it's not enough for something in the ground to be contrastive, in the F-marking sense, for a strategy to be evoked. There has to be a reason for it terms of discourse coherence.

So, in order to really account for the distribution and interpretation of FR accents, we need these information packaging notions above the focus semantic value (i.e kontrast/F-marking). This is exactly what alternative proposition type analyses lack. Once we accept that the bulk of what is going on here is to do with IS, we see the dominant prosodic signal we are dealing with is pitch accent placement rather than pitch accent shape. Moreover, while adding the rise seems necessary to evoke a strategy in some cases, e.g. when the response has the form of a direct one, it's not necessary in others. Either way, we maintain the same interpretation for the rise: it signals that the dialogue has not come to a stopping point.

This analysis has implications for development of higher level semantic and pragmatic models. We don't want to attribute pitch accent shapes to particular IS labels or discourse categories or, more generally, to analyze intonation components as operators on the proffered content of the utterance that carries them. Instead, they signal that a particular discourse configuration is in play. Sometimes the contribution of a rise parallels what is signalled by the other structural components of an utterance. When it doesn't, the hearer has to reconcile conflicting signals in a way that maintains discourse coherence. This may involve referencing higher levels of the dialogues task structure, or even more individual level characteristics of the other participants. We look at the effect of higher level contextual effects such as task, role and dialect in the next chapter.

Chapter 7

Higher Level Dialogue Factors: Task, Role, Dialect

7.1 Introduction

So far, we have examined the relationship between prosody and interpretation with respect to what we might call the micro-context: the utterance itself and the immediately dominating question under discussion. For example, in Chapter 4, we examined the effect of specific cue words, while in Chapter 5 we looked at Question-Answer Congruence (QAC) based constraints while fixing information structure. We saw that the experimental results of these studies were best understood via a hierarchically structured, task based view of discourse. In this chapter, we look beyond the local constraints of the QAC and the semantically proffered content to see what effects might come from the macro-context. One way to do this is to look at the distribution of rises in different tasks.

Our current analysis of terminal rises is that they reflect a speakers view of the discourse structure. In particular, they signal whether the discourse (i.e. task) has come to viable stopping point or not. Why should a speaker need to do this? One explanation is that such signals facilitate co-ordination between conversational participants. Taking this view we would expect such signals to become more prevalent as the actual need to co-ordinate is becomes higher – when speakers need to maintain a representation of the common ground of a certain quality. If this is the case, then we would expect more rises to occur in task-oriented dialogues than in conversational dialogue. Moreover, we should be able to see this as a reflection of the general structure of the dialogue.

To examine this, we look at the IViE (Intonational Variation in English) corpus

(Grabe, 2004). The corpus contains speech of various styles including isolated read sentences as well as spontaneous conversational and task-oriented dialogues (the map task) from speakers of urban regions of the United Kingdom. In the first part of this chapter we look at the boundary intonation of speakers from Cambridge (i.e. Standard Southern British English) in these different modes of speech. The motivation for looking at this dialect in particular is that, out of those included in the corpus, the intonation pattern for this region's declarative statements has been found to be the most pervasively falling or low at the boundary in read speech. Thus, rises are more likely to be seen as deviations from the canonical.[1] When we observe rises, we expect them to mean something more than just a phonological boundary. So, looking at this data enables us to look at the effects of task and role on the frequency of rises, as well as giving as a more general view on how task-oriented and conversational dialogue differ.

Task, role and move type are factors which are dependent on the actually unfolding discourse. However, we must also admit effects of discourse external factors on the distribution of prosodic forms. A salient feature captured specifically in the IViE corpus is, of course, dialect. The second part of this chapter we look at the variation in boundary prosody for different dialects. The goal of this last study is specifically to examine possible intonational variation in English with respect to continuous rather than discrete tonal category labels. From the existing literature we know that in Belfast English declarative statements and questions are rising ($\approx$ L*H%) by default (Cruttenden, 1997, Grabe, 2004). However, we would like to know whether there are differences between these broad move types beyond general shape. In investigating such questions, I will try to show that in viewing the data as continuously varying we can recover and extend the existing results without the labour of manual tone annotation.

This chapter is organized as follows. Section 7.2 reviews previous work on intonational variation and higher level discourse features. Section 7.3 presents the contents of the IViE corpus and describes the additional annotations and derived features that we will be examining in the studies to follow. Section 7.4 looks at the differences between the conversational and map task data in terms of distributions of dialogue moves. We indeed find differences in the distributions of moves, particularly in the rate of affirmatives is higher in the map task. With such structural differences in view, Section 7.5 looks at the boundary intonation of these data sets, described in terms of Legendre polynomial coefficients. Here we do find more rising features in the map task data, particular with respect to instruction moves. We don't find any strong connection between rising features and speaker changes or affirmatives. I argue that these results fit best with our current analysis of the meaning of terminal rises.

[1] It also makes it the most similar to the American English speech we have been looking at.

Section 7.6 examines dialectal differences in a read portion of the IViE data, and particular how the distribution of contour features relate to existing tonal annotations. The conclusions of these studies are summarized in Section 7.7

7.2 Background: Prosody and Higher Level Dialogue Features

As discussed in previously, a good number of works have investigated the relationship between prosody and lower level dialogue features such as individual dialogue acts/moves (cf. Chapter 2). From a recognition point of view, combining higher level dialogue (game) structure with an intonation model has been shown to improve word and move recognition on task-oriented speech (Taylor et al., 1998, Wright Hastie et al., 2002). Prosodic features have also been widely employed in, for example, topic segmentation (Shriberg et al., 2000, Tür et al., 2001) and detection of points of high involvement ('hot spots') (Wrede and Shriberg, 2003). So, there is clearly information to be gained about higher level discourse structure from prosodic features. In terms of situational factors, interest has recently moved towards various forms of role recognition in multiparty speech. For example, Salamin et al. (2010) report that a prosody based recognizer performs well above baseline using 5-fold crossvalidation on news bulletins and talk shows, although prosodic features do not appear add anything to a recognizer based on turn-taking features. Vinciarelli et al. (2011) investigate classification based on socio-emotional rather than task-based roles, e.g. protagonist, supporter, neutral, gatekeeper, attacker (Bales, 1969) in the AMI corpus (Renals et al., 2007).[2] Again, classifiers based on bi-gram type turn-taking models out perform those based on standard aggregate statistics for F0, energy and mean speech rate over turns. In this case, however, the combined model has the best results (59% accuracy over a 26% baseline, 49% for the turn bigram model). Examination of mean F0 and energy suggests that prosodic behaviour is conditioned by the previous role. However, it is not clear how these results generalize to other domains.

Recognition tasks like the ones discussed above generally do not investigate in any detail how prosody varies with different categories. So, the contribution of prosody to these different discourse dependent speaker characteristics is not well understood. However, some investigations of this type have been carried out from the perspective of improving expressive speech synthesis. Again analyzing data from the AMI corpus, Charfuelan and Schröder (2011) find that 'Assess' type moves are produced with more vocal effort (i.e. tenser voice quality).[3] In terms of role, they found project

[2] Augmented Party Interaction.

[3] Moves: Assess, Elicit, Suggest, Inform.

managers to have higher average F0 and vocal effort.[4] This extends Charfuelan et al. (2010) which finds that dominant participants in general exhibit 'louder' voice quality features.[5] These studies provide broad descriptions of specific roles, however they don't look at the contribution of intonation features like terminal rises in any detail.

The connection between dominance and prosody has, of course, come up in theoretical works on prosodic meaning. While accounts have been proposed linking submissiveness and rises (Merin and Bartels, 1997, Gussenhoven and Chen, 2000b), few works have pursued this as an empirical question. The few data-driven investigations in this area cast doubt on the connection between submissiveness and high boundary intonation. In a qualitative analysis of sorority speech, McLemore (1991) found that rises were used in by senior sorority members to take and hold the floor in monologues. That is, these speakers used rises in situations where they were dominant socially and as well as in terms of the speech task. However, rises were perceived as expressing uncertainty in narratives by uninitiated members of the group. Based on the observed distributions, McLemore argues that boundary tunes have a diagrammatically iconic meaning: rises are connecting, level tones are continuing, while falls are segmenting. However the actual interpretation of rises depends heavily of situational and cultural conventions. Within the sorority, rises were conventionally used to elicit involvement amongst members. Similarly, Cheng and Warren (2005) find rises to be more prevalent in discourses where one person has a socially dominant role, when comparing several different dialogue types, e.g. academic supervision versus informal conversation in an office. Moreover, it is the dominant participant who produces the rises. So, it seems that terminal rises associate with dominance, not submissiveness conditional on situational factors. Cheng and Warren do not conditionalize over different move types, so it's not clear whether the more one-sided conversations simply involve, for example, more questions. Both of these studies abstract away from the phonetic detail using categorical tonal analysis. So, it seems a detailed, quantitative look at the factors in play is still wanting.

Nevertheless, these empirical investigations clearly indicate that one of the factors that governs the distribution of rising features is the type of dialogue. Work on speech style has generally focused on the differences between read and spontaneous speech. The trend is for such investigations to find more intonational variation in the latter. For example, Hirschberg (2000) report more rises in spontaneous speech than in read speech (71% versus 93% for declaratives) comparing data from the *atis0*

[4]Roles: Roles in the AMI corpus : Project Manager, Marketing Expert, User Interface designer, Industrial designer.

[5]Dominance annotations on the AMI corpus (Jayagopi et al., 2009, Aran et al., 2010) were done naively: i.e. annotators were told to rate relative dominance but where not given a prior definition of dominance, nor were they told to attend to any specific speech features.

corpus. Similarly, Kowtko (1997) found the pattern of rises and falls of one word turns to be clearer in read versions of HCRC map task dialogues than in the original spontaneous versions, with respect to specific move types (e.g. 92% versus 72% rises for ALIGN moves). Direct comparisons of conversational and task-oriented speech have mainly focused on the greater need for affectual/emotional modelling in the former (Campbell, 2005, Wilks et al., 2011). However, more concrete differences can be seen in the diverging results about the role of prosody in different task domains: Benus et al. (2007) characterizes backchannels in American English as having rising pitch and higher intensity based on data from the Columbia Games Corpus, while Truong and Heylen (2010) do not find this in the AMI corpus of multiparty meetings. However, I have not found any works directly comparing spontaneous dialogues of these types in terms of prosody. In general, studies of speech style have not really attempted to situate their findings in a larger theory of prosodic meaning or discourse structure, or vice versa.

Dialectal intonational variation has received more attention from linguistic circles than situation induced differences like task/role. Much of this work has attended to variation in tonal inventories. For example, Arvaniti et al. (2007) argue that American English speakers from Minnesota lack H*/L+H* distinction, which is present in Southern Californian but where such peak delay seems to predominantly signal of emphasis. With respect to boundary intonation, the use of final rises on statements has been of considerable interest to descriptions of region differences in spoken English. Such rises have been associated with the English spoken in Australia (Guy et al., 1986) and New Zealand (Britain, 1992), as well as various British cities outside of South England, e.g. Newcastle/Tyneside (Local et al., 1986), Liverpool (Knowles, 1978), Glasgow (Brown et al., 1980, Mayo et al., 1997), and Belfast (Jarman and Cruttenden, 1976). Cruttenden (1997, pg. 133) notes that these statement rises come in different forms, with a rise proper preferred in Glasgow, a rise-fall preferred in Wales, but a rise-plateau preferred in Belfast, Liverpool, Newcastle and Birmingham. Following the impressionistic analysis of Liverpool intonation of Knowles (1978), Cruttenden suggests that apparent differences of default rise/fall in English could just reflect differences in phonetic implementation if we treat rise-plateaux as truncated falls. Grabe et al. (2000) study truncation and compression in nuclear falls and rises in read nonce words. Their data indicate that while Cambridge, Leeds, Newcastle speakers produce the same tunes for statements and questions (i.e. falls and rises), Cambridge and Newcastle speakers compress these tunes, while Belfast and Leeds truncate. Given this is division, it is unclear that we can attribute the rising features of the urban north to truncation. However, this highlights potential differences between underlying phonological variables and phonetic implementation.

As noted in Grabe (2004), work on intonational variation has for the most part

focused on cross-language comparison of single dialects. However, there has been a general dearth of information on cross-speaker, cross-dialectal variation. This mirrors this situation with respect to intonational variation in speaking style and similar higher level discourse features. As noted above, there haven't really been systematic, quantitative comparisons of different styles of spontaneous dialogue focusing on intonation. However, previous studies suggest that we do need to take discourse situational and dialectal differences seriously if we are to develop a linguistic theory of intonation. For example, Cruttenden (2007) found the read speech of one Glasgow English speaker to exhibit characteristic RP falls, while their conversational speech exhibited final rises expected for the urban north pattern. The IViE Corpus (Grabe, 2002) was created to fill some of these gaps. However, thus far only results from the read portion of the corpus have been reported. So, in the following sections we will investigate the higher level features of a dialogues, e.g. role and goals, based on the IViE spontaneous dialogue data. We will return to the specific findings about regional differences in the IViE read data in Section 7.6. But first, we briefly review the contents of the IViE corpus.

7.3 Data

7.3.1 The IViE Corpus

The IViE corpus was developed to study differences across regions, speakers, and styles in a systematic way (Grabe, 2004). The corpus contains data from several regions of the British isles: London, Cambridge, Leeds, Bradford, Newcastle, Belfast, Dublin, Liverpool and Wales. Although Leeds and Bradford are very close geographically, the speakers come from different language backgrounds: the Bradford speakers were Punjabi/English bilinguals. Similarly, the London speakers were of Caribbean descent and spoke West Indian English. This means that the only region sampling Standard Southern British English was Cambridge. Twelve speakers (6 male, 6 female) from each region were recorded between 1997-2000. The speakers were 16 years old at the time of recording and had been born in and grown up in the region. The recording involved 5 tasks:

- Reading of isolated utterances: 8 declarative statements (dec), 3 declarative questions (dqu), 3 polar questions (yno), 3 wh-questions (whq), 5 alternative questions (coo).
- Reading of the Cinderella story (rea).
- Retelling of the story guided by pictures (ret).

- Map task dialogue (map): Each participant was given a map of a small town. Participants took one of two roles: Instruction giver and follower. The instruction giver's map included a route around the town. The goal was for the giver to explain the route to the follower, who had to trace out the route on their own map. The task ended when the route was completed to the satisfaction of both participants. Maps differed in place names and locations of landmarks, so speakers had to work to establish common ground. Speakers were separated by screens so they could not see each other. More details about this task can be found in Anderson et al. (1991).
- Free conversational dialogue (conv): Participants discussed smoking. Unlike the map task this was a face-to-face dialogue.

IViE Tone Transcription

The IViE intonational transcription system is a variant of the ToBI system. The goal was, however, to create a system for which meaningful cross-dialect comparisons could be made. The phonological level involved annotation of ToBI style accent and boundary tones. It is about this tier that results have been reported.[6] As is usual for this sort of tone labelling, annotators were able to look at the F0 curve as well as listen to the production during the process. To try to maintain comparability, utterances were 'not labelled in isolation, but in comparison with a directly comparable utterance produced by another speaker from the same dialect' (Grabe, 2004). Table 7.1 reproduces descriptions of the tones used. Note, the tone inventory includes a level boundary tone not included in the American English ToBI set. This is to accommodate rise plateau patterns attested in, for example, English of Northern Ireland and Glasgow (See discussion in, of example, Ladd (2008)).

Transcriptions were done for 7 of the regions (*not* Liverpool or Wales). Within each region tonal annotations were provided for half of the speakers for the isolated read utterances. Transcription for the other tasks was done for one male and one female speaker. Since there is only a small amount of labelled data for the dialogues, we will not really be using the tone labels in the first part of our study. However, we will look more at the relationship between tone labels and F0 features in the second part of this study, which focuses on the read utterances (Section 7.6).

[6]The transcription also includes annotation of prominence location and 'phonetic' transcription of pitch patterns around prominent syllables. This is to account for difference in phonetic implementation. See Grabe et al. (2001) for further details.

Accent	Description
H*L	High target on prominent syllable followed by low target
H*	High target, common in initial position in so-called flat hats
!H*L	Downstepped high target, low target
L*HL	IP internal or IP final rise-fall: Low target on prominent syllable, high target on next syllable followed by low target
L*H	Low target on prominent syllable followed by high target
L*	Low target
H*LH	IP internal or IP final fall-rise: high target on strong syllable, low, high
Boundary	
H%	high target
%	no pitch movement at boundary
L%	low target

Table 7.1: Accent Tones used in the IViE transcription Grabe (2004). Note: not all of these tones appear in the reported results.

7.3.2 Additional Annotations and Features

Dialogue

In Sections 7.4 and 7.5 we will look at differences in the conversational and map task dialogues produced by the speakers from Cambridge. Out of the official IViE release, only short portions from two pairs of speakers were transcribed and annotated for sentence type level for each of these dialogue types, so additional annotation was undertaken to get a better view of the data. Since we are primarily interested in utterance boundary prosody, I manually segmented the data into utterances. The original goal was to mark utterances which would correspond to whole meaning units rather than phonological phrases (cf. Swerts and Geluykens (1994)). The reasoning was that this would be a more conservative measure of the frequency of rises. Sentential segmentation, for example, delimited whole propositions including any embedding. Similarly, imperatives utterances basically mapped to one action (i.e. one segment of the route). A number of sub-sentential clauses also formed separate utterances, e.g. an NP or VP as an answer or a modifier separated by an affirmative, which were as XPs. Utterances were labelled for sentence (syntactic) and move type:

- Sentence type: Declarative (dec), Imperative (imp), Polar question (yno), Wh-question (whq), Tag (tag) question, Affirmative (affirm), Negative (neg), Cue word (cw), If antecedent (IFA), XP (XP).

- Broad dialogue moves: Affirm, Neg, Contra (direct contradictions),[7] CW (cue words), Inform,[8] Instruct, Q (non-syntactically marked question), YNQ (polar question), WhQ, (wh-question), Tag (Tag-question), sync (synchronize/align).

The labels given here were chosen to be quite shallow reflections of dialogue moves. Obviously, in many cases, one sentence type dominated a move type (e.g. the wh-questions). The rationale for using such broad types was to keep the annotation in terms of easily identifiable categories as a baseline for more detailed structural analyses to be attempted in the future. The main points of variation were, of course, with declaratives which we see as instructions, informing moves, and questions, amongst other moves. Similarly, instructions came in many forms. The *sync* category captured utterances in the map task like 'You should be at the Anne's Arms' which were not quite questions, instructions or Inform moves. The few cue words that were not affirmatives (CW) were mostly *really*s and *oh*s and the like. In the investigations to follow we will concentrate on the most populous and easy to identify categories: Affirmative, Instructions and Inform moves.[9]

Note, I did not include a backchannel category because it was not clear to me that the distinction could be reliably made. Certainly it did not seem that it could really be said of any of the affirmatives in the map task that they were just signals of attention. The question types were kept separate since we assume that they induce different types of structures/expectations in the dialogue (Kowtko et al., 1991, Ginzburg, 2012, Farkas and Bruce, 2010). Following Section 5.4.2, the target area for analysis was the stretch of speech from the last prominence rather than the last word. Extension away from the last word was generally due simply to stress assignment in compounds (e.g. *bowling* alley) or deaccenting of pronouns, (e.g. *about* it). Utterances with speaker overlap at the target were excluded from the prosodic analysis. However, placement of overlapping affirmatives was noted since we are interested in how the distribution of these differ in each type of dialogue. Fragments and filled pauses were also noted, though a distinction wasn't made between the two and they do not figure in the prosodic analysis either. The role of the speaker of each utterance was also noted. There are two distinct roles defined in the map task: the instruction giver (1) and the follower (2). In the free conversation we assume that speakers have the same role, which is simply a participant (3) since the goal of this task was to simply talk about smoking. Note, in this way the role encodes the overall dialogue task.

[7]This probably should have been rolled into the Inform class, but there aren't that many of them.

[8]This is basically the 'statement' or 'explain' move. I call it 'Inform' after the AMI corpus DA set, simply because it seems to describe what the move does most accurately.

[9]For reasons of time, I couldn't include detailed response type information of the kind we looked at in Chapter 5, even though this is clearly important for intonation. Of course, this is something for future work.

Read Utterances

Isolated read utterances from the corpus were also examined. Timing data was automatically extracted using the Penn Phonetics Lab Forced Aligner and the production stimulus list. Word boundaries were then manually corrected by the author. Since these utterances were produced out of the blue, we take them to have broad focus. Thus, the target segment was taken to be last word. Moreover, since there is no contextual reason to assume otherwise, we associate the 'standard' move type with each of the sentence types, e.g. we assume a declarative statement (dec) to be played as an Inform move etc. The exceptions, of course, are the declarative questions.

7.3.3 F0 Features

The target segments speech were extracted based on the boundary annotations. The F0 contour data was extracted using the autocorrelation method as implemented in Praat. Pitch range settings were automatically determined using the method described in Evanini and Lai (2010). Utterances which produced less than 5 F0 points were discarded. The F0 data was normalized into semitones relative to the median F0 value (Hertz) for each speaker. In the dialogue study, these medians were calculated over the combined data from the dialogues and read utterance for the Cambridge speakers. For the dialect study we normalize only with respect to the read speech. F0 contours were then smoothed using a Butterworth filter. The contours were then approximated using Legendre polynomial decomposition of order 4 as in Section 5.4.2.

7.4 Distributions of Moves in the Dialogues

In this section, we look at the distributions of moves in the conversational and map task dialogues. This will give us a general view into their discourse structure which will help us relate the intonational data to the different dialogue types later on. Besides the frequency of move types, we would like to know if the type and rate of affirmatives is the same in the different dialogue types. We expect the default move after an Inform or Instruct move to be an affirmative, so we would like to know how often this actually happens. Similarly, we would like to look at how often speaker changes happen, and what sort of move takes place at these sorts of junctures.

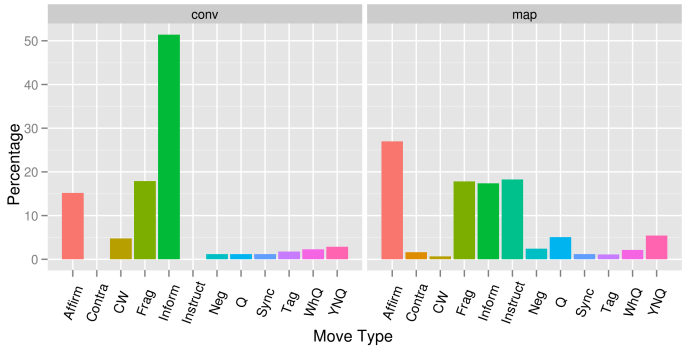


Figure 7.1: Proportions of moves for Cambridge speakers: $N_{conv} = 430$, $N_{map} = 1287$.

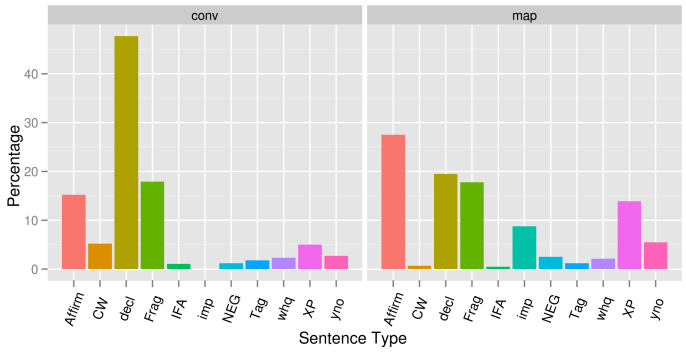


Figure 7.2: Sentence Types for Cambridge speakers

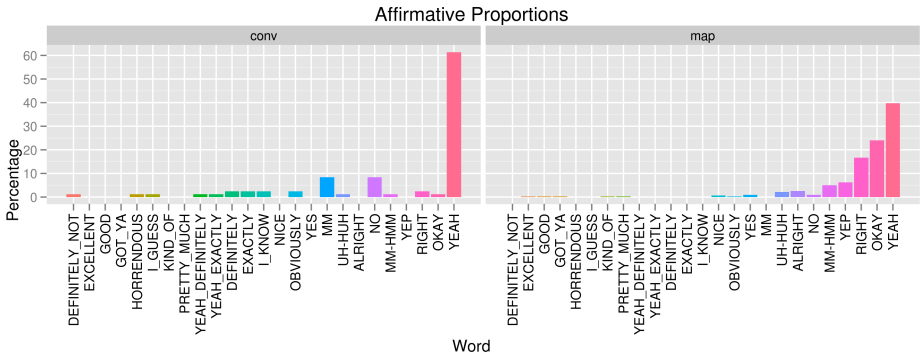


Figure 7.3: Affirmatives (Cambridge)

7.4.1 Type Frequencies

Utterance segmentation of the dialogues resulted in 430 and 1287 utterances for the conversational and map task sets respectively. Figure 7.1 show the proportions of each move type for the two different types of dialogue (including fragments/filled pauses). Out of the turns represented in the graph, approximately 3% (conv) and 4% (map) were overlapping at the target. Of the affirmatives 5% and 2% (resp.) were overlaps. We can immediately see some distributional differences between the dialogue types. The most obvious difference is, of course, that the map task involves Instruct moves while the free conversation does not. Beyond this we see that a higher proportion of moves are affirmatives for the map task. This fits with the idea that the task-oriented dialogue requires more explicit co-ordination. Similarly, we see a higher proportion of question type moves in the map task. In Figure 7.2 shows that the main difference in terms of sentence types is that we see a lot more (non-fragment) non-sentential utterances in the map task, which is to be expected given that this task involves the identification of more objects. As we might expect, not all Instruct moves were syntactic imperatives: declaratives, polar questions, and non-sentential utterances like if-antecedents were used as instructions.

Figure 7.3 shows the frequencies of specific affirmatives. The first thing to note is that *yeah* does not dominate the distribution as much in the map task. Instead we see a higher percentage for' *okay*. This mirrors affirmative frequencies for other task oriented data sets like the Columbia Games Corpus (Gravano, 2009). The greater frequency of *okay* makes sense when we consider that this affirmative signals acceptance rather than agreement (cf. Section 4.3.4). In the map task, we see *okay* being used to accept a task (from the follower side), or signalling that a speaker is ready to move on (usually the giver), i.e. accepting the current state of the game. In any case, we can take a relatively high proportion of *okay*s as an indicator of how task-oriented a dialogue is. Similarly, the increased frequency of *right* is appropriate in the map task

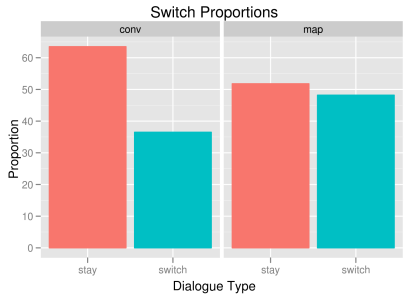


Figure 7.4: Proportion of speaker changes for both dialogue types

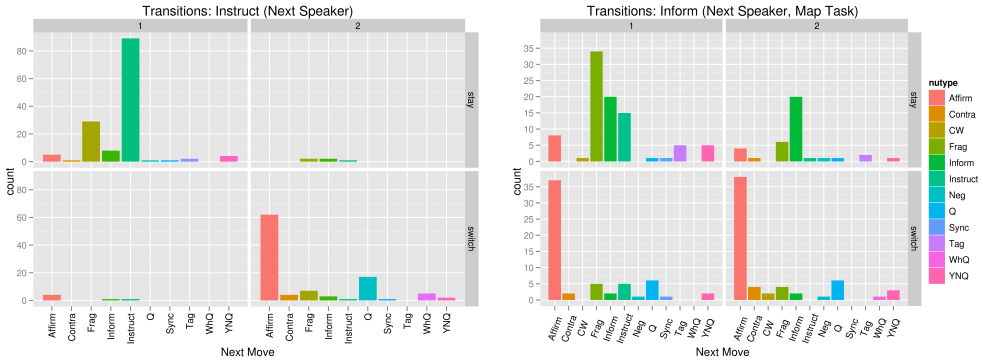


Figure 7.5: Speaker changes: Moves following Instructions by next speaker, i.e. the upper left facet shows 1-1 transitions, the upper right 1-2, where 1 = giver, 2 = follower.

because this does require a high level of alignment/agreement. Personal preferences aside, there is less actual need to put forth strong agreement in a random conversation in which the goal is to find out what other people think, rather than to come to a resolution.

7.4.2 Speaker Changes

In terms of turn-taking, we can observe that there were more speaker switches in the map task dialogues (Figure 7.4). This seems to reflect the task structure of the dialogues. Figure 7.5 shows moves following instruction for the map task, broken down by role and transition type. Here we see, an instruction is likely to be followed by another instruction by the giver or else an affirmative by the follower. Unsurprisingly, the vast majority of instructions were issued by the speaker in the instruction giver role. However, this sort of role distinction doesn't hold for Inform moves. In information exchange type subtasks (games), where the participants are really determining the common ground rather than building the route, the roles are more even.

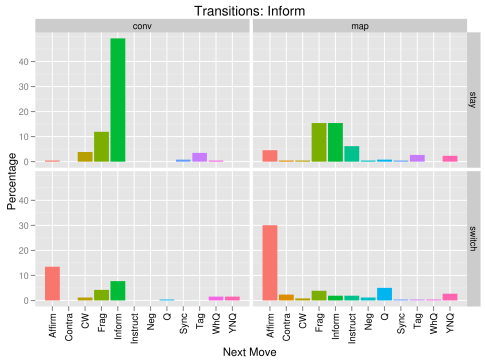


Figure 7.6: Speaker changes: Moves following Informs grouped by next speaker

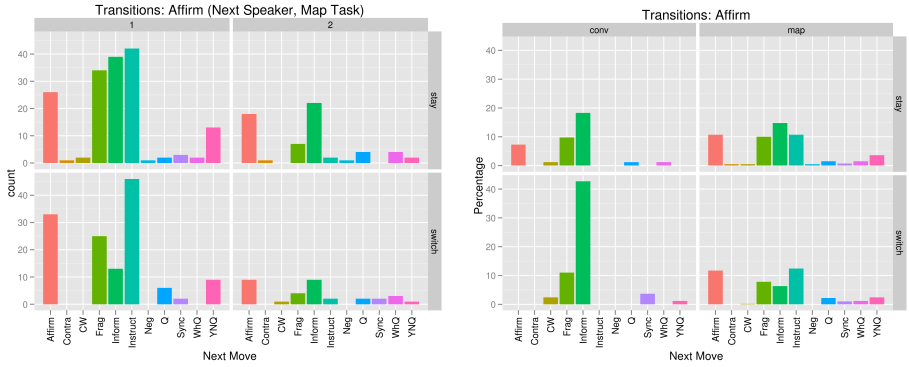


Figure 7.7: Moves after affirmatives, by role.

Comparing this with transitions after Inform moves in conversational speech (Figure 7.6), we see a striking decrease in the proportion of switches to affirmatives. This is balanced by an increase in cases where the speaker retains the floor producing another Inform move. So, it seems that conversational speech is characterized by longer stretches where one speaker holds the floor. This suggests that describing ones opinion takes more utterances and but also induces less explicit affirmation from the hearer. However, part of the reduction in affirmatives could also be attributed to the face-to-face set up of the conversational dialogue. In this case, speakers could be using other channels to signal feedback. That is, there is even greater need to co-ordinate through the speech signal in the map task.

Figure 7.7 shows moves after Affirm moves. In the map task we see a number of Affirm moves replied to with another Affirm by the other speaker. We don't see this for the conversational dialogue, though we do see affirmative doubling from the same speaker. Switches are slightly more frequent after affirmatives for the conversational data. For the map task, instruction givers are more likely to retain the floor after producing an affirmative, while followers are more likely to give it up (Figure 7.8).

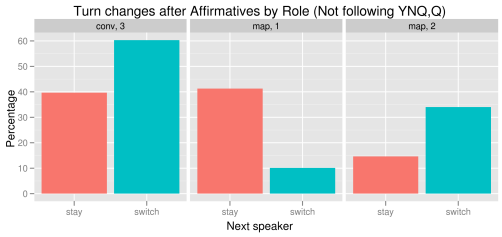


Figure 7.8: Speaker changes after affirmatives

7.4.3 Summary: Move Characteristics of Task Oriented and Conversational Speech

Looking at the broad distribution of moves and turn changes we see that the task oriented and conversational speech exhibit clear differences. Especially salient is the fact that the task oriented speech is peppered with more affirmatives, particularly *okay*. There are more switches in this style of dialogue. This makes sense since the map task is characterized by sequences of instructions. If these action sequences go well the follower's contentful input is minimal – mostly they just provide affirmations. When the difficulties in this arise, the participants move to common ground building mode (Inform moves/). In contrast, speakers in conversational dialogues are more likely to make multiple Inform moves in sequence without feedback. This suggests less common ground building than making private beliefs public, i.e. expressing opinions which don't really require joint commitment.

7.5 Boundary Intonation in the Dialogues

7.5.1 Potential Connections between Move Structure and Intonation

In the previous chapter I argued that terminal rises provide a signal about the speakers expectations about what should come next, i.e. not a stopping point). In terms of turn-taking, this could go either way. In fact, this is the direction pointed to by other work which suggests that plateaux at the boundary to be turn holding (**?**Gravano and Hirschberg, 2009). So, it would be somewhat surprising to find a strong connection between boundary intonation and speaker changes. Nevertheless, if we were to think that rises, or a similar prosodic marker, contributes some sort of turn keeping signal, then we would expect to see it more in the conversational data, since we see relatively more stays. If we expect that they are turn-giving we would expect to see more rising features in the map-task data.

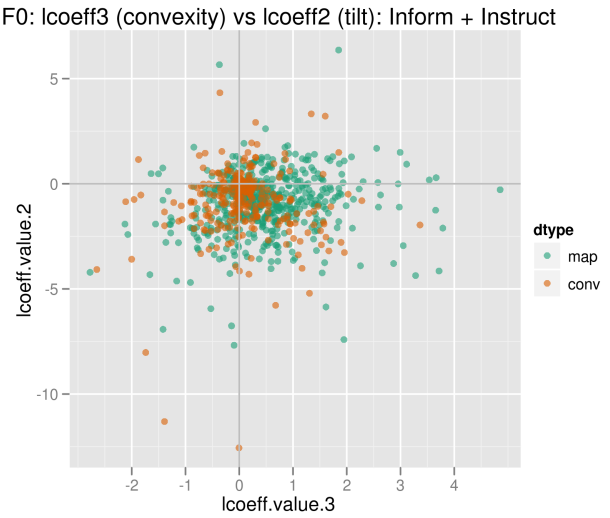


Figure 7.9: Inform and Instruct moves: Tilt and convexity.

The higher frequency of affirmatives in the map task dialogues supports the idea that task-oriented dialogues require greater synchronization and co-ordination. This is especially the case for the map task as the speakers realize that their maps are not actually the same. Success in conversational speech is less contingent on carrying out individual actions in general. In this vein, we hypothesize that speakers are more likely to signal that 'we're not done yet', via rises, in the map task. Conversely, we'd expect less rises when diversion from the task at hand is allowed, i.e. in free conversation. We would not expect this to be reflected by whether or not a speaker change takes place.

So, we want to look at the distribution of intonational features and how this relates to speaker switches. From the previous section we know that the majority of utterances are Affirm, Inform or Instruct moves. These are non-checking moves, so we expect that if they have rises they will be more like fall-rises (low rises) rather than the convex nuclear tones we saw associated with declarative questions.[10] However, instead of trying to make a categorical judgement about shape, we will instead look at distributions of Legendre polynomial coefficients, particularly those representing tilt (LC2) and convexity (LC3).

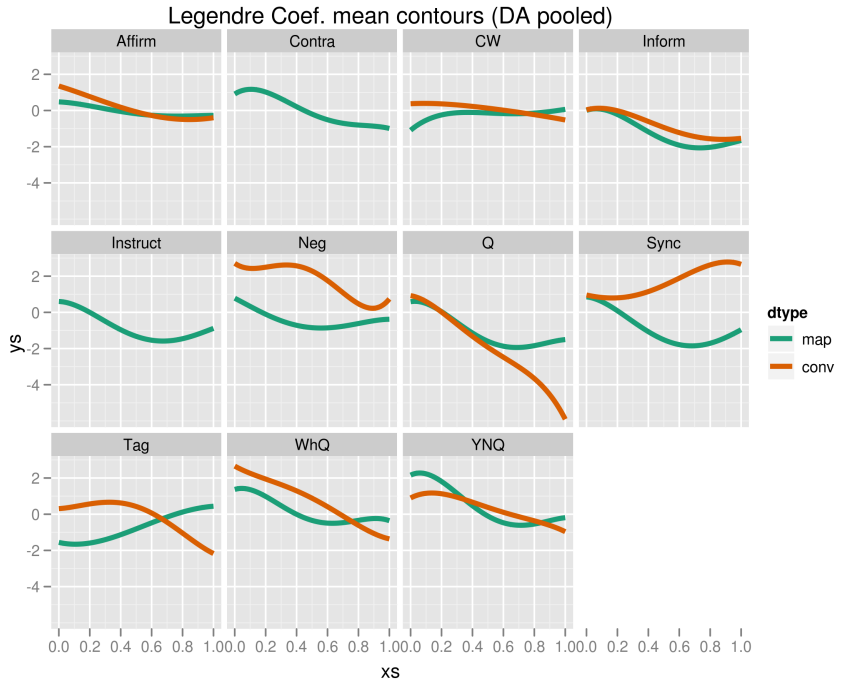


Figure 7.10: Mean contours based on Legendre coefficients.

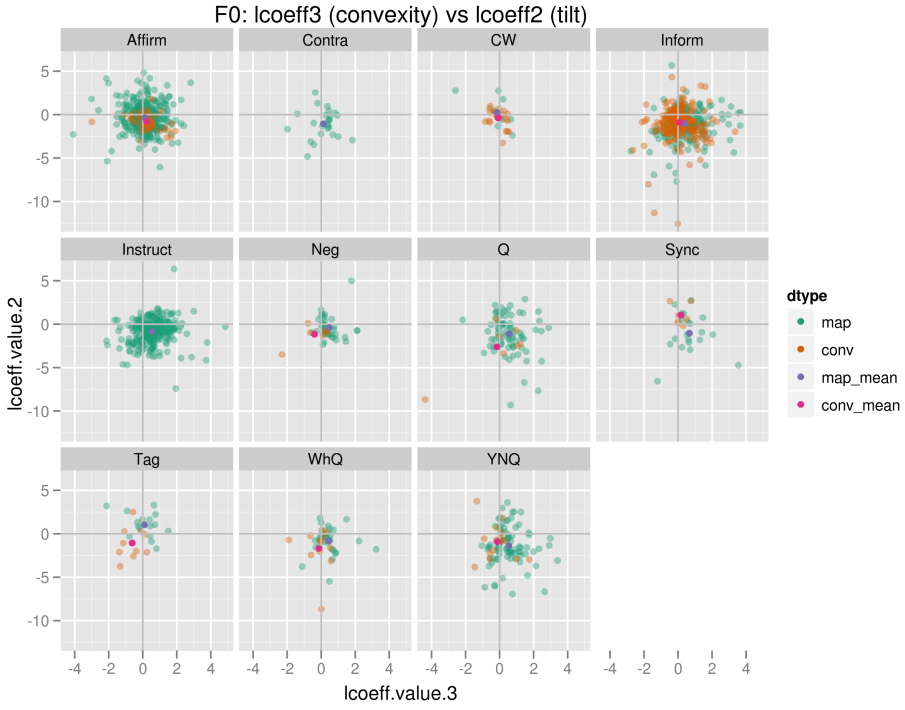


Figure 7.11: Tilt (LC2) and Convexity (LC3), by move type, with means.

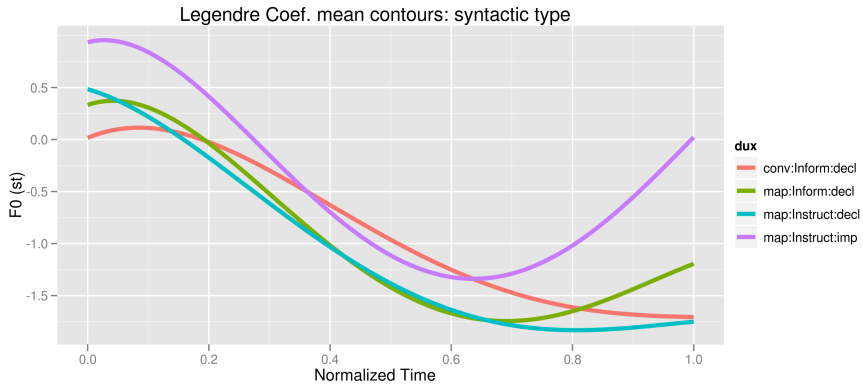


Figure 7.12: Mean contours Instruct and Inform moves based on Legendre coefficients grouped by syntactic type.

7.5.2 Exploration of Intonational Features

Are there more rises in the map task dialogue? Figure 7.9 compares the distributions for Inform and Instruct moves with respect to the derived coefficients for the 2nd (tilt) and 3rd (convexity) Legendre Polynomials. Informs and Instructs make up 44% (262+284/1235) of moves in the map task, while Inform moves make up 66% (270/408) of the utterances in the conversational speech. These moves provide most of the 'new' content in the dialogue. Moreover, we would assume both moves types to be canonically falling (Cruttenden, 1997). So, we take this subset of moves as representative of whether task affects the distribution of rising features. The graph shows that the map task data extend further along positive side of x-axis (LC3) indicating greater convexity. The contours derived from the mean values of each of the 5 Legendre polynomial coefficients (Figure 7.10), show the low rise tendency of map task utterances. In particular we see that, on average, Inform and Instruct utterances present a fall-rise sort of shape. Interestingly we can also note, that the declarative questions are falling. Figure 7.11 shows the distribution of values for LC2 ad LC3 for the rest of the move types. Breaking things down in this way, we see that Instruct moves are mostly situated in the positive LC3 space. The space taken up by the Inform moves is more or less completely overlapped for both dialogue types. So, more rises come from the instructions in the map task. The distribution of affirmatives in the map task extends into the positive LC2 space, indicating rising tilt. So, overall it does seem that map task data can be characterized as having more rising features.

Do we see differences controlling for syntactic type? Figure 7.12 shows mean con-

[10] Although this was a generalization for SAE speakers, we will see shortly that this holds for the Cambridge speakers too.

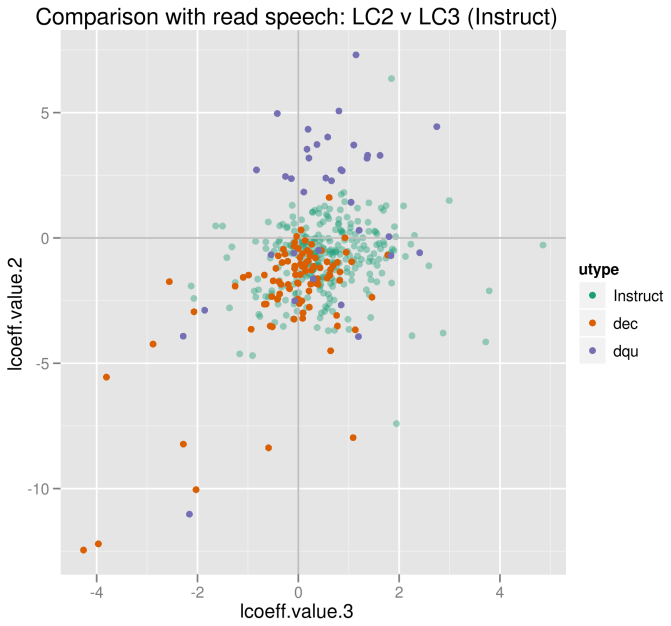


Figure 7.13: Instructions overlaid with read declaratives

tours for Inform and Instruct moves which are declaratives, as well as the imperatives for comparison. Again we see that, on average, Inform declaratives are less convex in the conversational speech than in the map task. However, it also seems that map task imperatives are more rising than declarative instructions. So, within the Instruct moves, syntactic imperatives are particularly particularly rising.

As mentioned previously, we assumed that both Inform and Instruct moves were canonically falling at the utterance boundary (as opposed to fall-rise). But are they in fact more rising than isolated read Inform moves (i.e. declarative statements)? To look at this we can compare the shape of these features with those of declaratives read in isolation. Figure 7.13 shows LC2 and LC3 for the Instruct utterances overlayed with coefficients from declarative statements (Inform) and questions read in isolation.[11] From this we can see that a good mass of the Instruct utterances fall in the region between the read statements and questions, i.e. fall-rise territory. The population of Inform moves in this region is not so great (Figure 7.14), although we do see more utterances with positive tilt. While there is complete overlap with the statements, neither move type goes into the territory of the declarative questions (convex with large positive tilt). It is also interesting to note that while the isolated declarative

[11]Unfortunately, there are no imperatives (or declaratives which could be construed as Instructions) in the read component of the IViE corpus. We will look more at the read speech in IViE in the second part of this chapter.

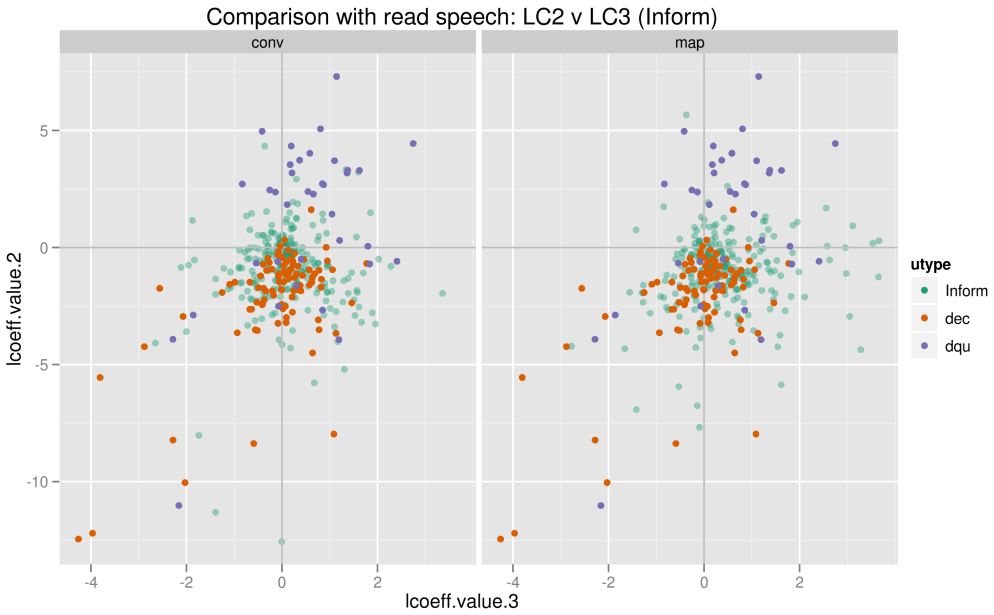


Figure 7.14: Inform utterances overlaid with read declaratives

question are characterized by large rises, the declarative questions in the dialogue are often falling (cf. Figure 7.11).

7.5.3 The Effects of Role, Move and Syntactic Type

One of our main goals was to get a better understanding of the relationship between higher level contextual factors and intonation. The basic point from the data exploration is that the utterances from the map task dialogues, particularly instructions, have a more convex shape than utterances from the conversational data. In order to quantify this we model the relationship between Legendre polynomial coefficients and discourse factors: role, move, and syntactic type. To do this, (non-nested) multilevel linear models were fit predicting the derived coefficients for the first three Legendre Polynomials, i.e. y_i represents observed height (LC1), tilt (LC2), convexity (LC3). The parameters of the model were as follows:

$$(1) \qquad y_i \sim N(\mu + \alpha^{\text{role}}_{j[i]} + \alpha^{\text{spk}}_{k[i]} + \alpha^{\text{move.syn}}_{l[i]}, \sigma^2_y), \text{ for } i = 1, \ldots, 1520$$

With group level coefficients modelled as follows:

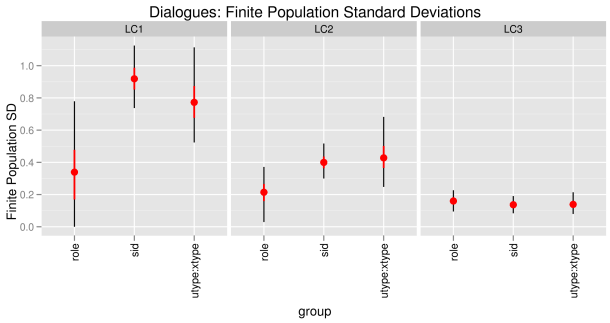


Figure 7.15: Finite Population Standard Deviations: Medians 95% and 50% intervals

$$\begin{aligned}
\alpha_j^{\text{role}} &\sim N(0, \sigma^2_{\text{role}}) \text{ for } j = 1, \ldots, 3 \\
\alpha_k^{\text{spk}} &\sim N(0, \sigma^2_{\text{spk}}) \text{ for } k = 1, \ldots, 12 \\
\alpha_l^{\text{move.syn}} &\sim N(0, \sigma^2_{\text{move.syn}}) \text{ for } l = 1, \ldots, 26
\end{aligned}$$

Fragments were excluded from the data. Note that not all move/syn combinations were present in the data. The model parameters were estimated using the package `lmer` in `R`. Figure 7.15 shows the estimated finite population standard deviations for the predictor groups for each of the LC models.[12] Most of the variation in the LC1 model comes from the speakers themselves and the move type, while the estimated standard deviation for the role group is not significantly greater than zero (similarly for LC2). The role estimate is significantly greater than zero for LC3, although the variance overall is smaller for this model.

Figures 7.16, 7.17, and 7.18 show the parameter estimates. We take estimates that are more than twice their standard deviation away from zero to be significant. In each case we see significant difference between speakers), in both positive and negative directions. When we look at convexity (LC3) we see significant effects of role. Being an Instruction Giver (1) increases convexity, while simply being a conversational participant (3) decreases it. In terms of moves, we see a positive effect on convexity for imperatives, in-line our previous observations. As we would expect, declarative Inform moves have lower overall height (LC1), as do the non-sentential Inform and Instruct moves. The effect of Affirm moves was negative on convexity, but positive on height and tilt. With the estimated value for the instruction follower being positive, this again points to there being more rising affirmatives in the task-oriented dialogue. Yes/no and declarative questions have a negative relationship with

[12]The finite population standard deviation is a measure of the variance in the actual parameter estimates. Looking at these estimates basically a multilevel version of classical analysis of variance.

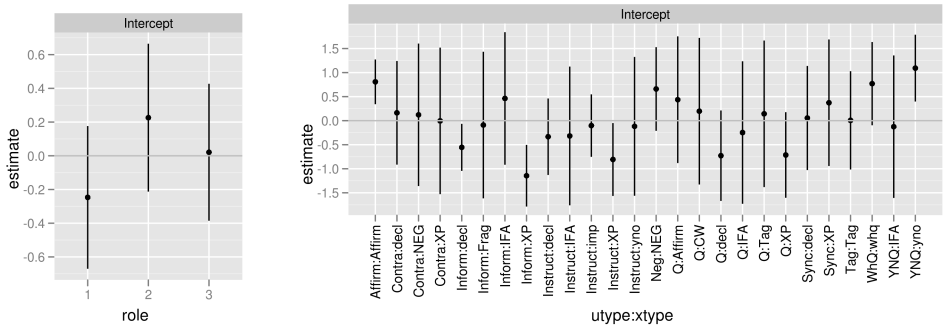


Figure 7.16: F0 height (LC1) parameter estimates.

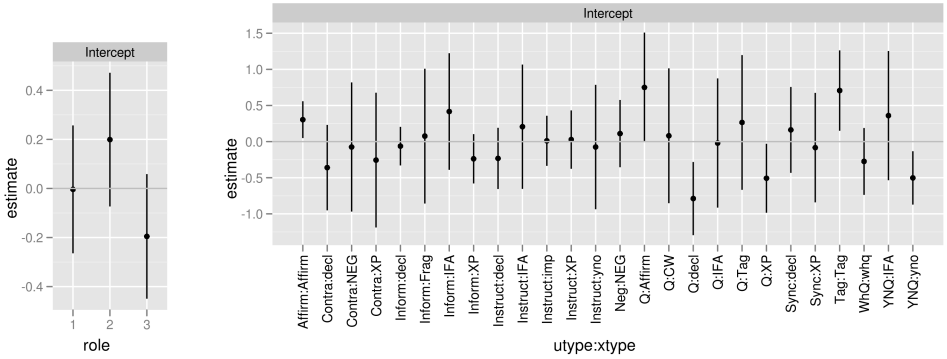


Figure 7.17: F0 tilt (lcoeff.value.2) parameter estimates.

tilt. This leaves us with more rising features on canonically falling imperatives and more falling features on the canonically rising declarative questions. This suggest that the specific questioning use of rises is less in play in these sorts of dialogues than the general connecting aspect.

We should note that the effects we have been through, though looking significant, are quite small. Based on the our analyses in other chapters, it is not hard to believe that there are many more factors at play. In particular, we expect the speaker engagement (beyond what is necessary to complete the task) to have an effect of the contour shape. Nevertheless, at this point it does seem to be the case that the boundary intonation of the map task data inhabits more of the rising pitch feature space. Instruction givers seem to produce contours of greater convexity. We see rising in two shapes: low convex rises which are prevalent for Instructions (particularly syntactic imperatives), and both convex and non-convex rises on affirmatives.

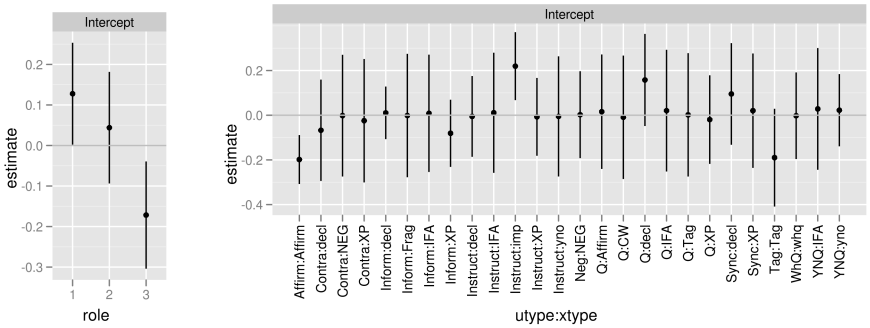


Figure 7.18: F0 convexity (lcoeff.value.3) parameter estimates.

Parameter	Std. Dev.
move:role	0.48
spk	0.42
move	0.90
role	0.82

Table 7.2: Estimated standard deviation σ for group level parameters.

7.5.4 Turn-taking

We now check to see if the rises associated with the task oriented dialogue are to be interpreted as turn holding or turn giving. To do this, we fit parameters for multilevel logistic regression models (stay=0, switch=1). First we look at non-intonational effects:[13]

$$(2) \qquad \Pr(y_i = 1) = \text{logit}^{-1}(\beta^0 + \alpha^{\text{role}}_{j[i]} + \alpha^{\text{spk}}_{k[i]} + \alpha^{\text{move}}_{l[i]} + \alpha^{\text{move.role}}_{s[i]}).$$

With group level coefficients modelled as follows:

$$\begin{aligned}
\alpha^{\text{role}}_j &\sim N(0, \sigma^2_{\text{role}}) \text{ for } j = 1, \ldots, 3 \\
\alpha^{\text{spk}}_k &\sim N(0, \sigma^2_{\text{spk}}) \text{ for } k = 1, \ldots, 12 \\
\alpha^{\text{move}}_l &\sim N(0, \sigma^2_{\text{move}}) \text{ for } l = 1, \ldots, 11 \\
\alpha^{\text{move.role}}_s &\sim N(0, \sigma^2_{\text{move.role}}) \text{ for } s = 1, \ldots,
\end{aligned}$$

Parameters were again estimated using `lmer` in `R`. The parameter estimates for the model (2) are shown in Figure 7.19, while Table 7.2 shows the standard deviation, σ,

[13]Note: This is not improved by including the syntactic predictors (Likelihood Ratio Test, $p > 0.9$, DIC is the same for both models).

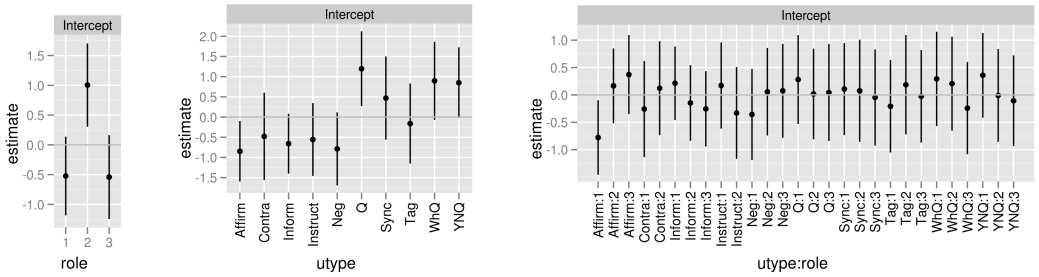


Figure 7.19: Speaker change (ts) parameter estimates

estimates for the groups. Roughly, we can estimate that the probability of switching differed by approximately ±20% by role, ±22% by move, and ±10% by speakers, and similarly the move-role interaction was estimated at about ±12%.[14]

Looking at the actual estimates in Figure 7.19, we see quite a large effect of being the instruction follower (role 2). Having this role increases the probability of switching by 25%. The trend for the other two roles is to hold the floor. In terms of move type, we see that the broad class of question moves increase the probability of switching, while content adding moves decrease it. Looking at the move-role interaction we see that affirmatives produced by the instruction giver are likely to result in stays. This reflects our previous observation that Instruction givers use affirmatives as a ready signal. The effects of the other move-role combinations are quite small in comparison.

What do intonation features add beyond this? We investigate this by adding Legendre coefficients to the predictors to the model. Let $\mathrm{lc}_{k,i}$ be the observation of the k^{th} Legendre polynomial coefficient for the i^{th} data point. Our model is then:

$$\text{(3)} \qquad \mathrm{lcpred}_i = \beta^{\mathrm{lc}_1} \cdot \mathrm{lc}_{1,i} + \beta^{\mathrm{lc}_2} \cdot \mathrm{lc}_{2,i} + \beta^{\mathrm{lc}_3} \cdot \mathrm{lc}_{3,i} + \beta^{\mathrm{lc}_4} \cdot \mathrm{lc}_{4,i} + \beta^{\mathrm{lc}_5} \cdot \mathrm{lc}_{5,i}$$

$$\text{(4)} \qquad \Pr(y_i = 1) = \mathrm{logit}^{-1}(\beta^0 + \mathrm{lcpred}_i + \alpha^{\mathrm{role}}_{j[i]} + \alpha^{\mathrm{spk}}_{k[i]} + \alpha^{\mathrm{move}}_{l[i]} + \alpha^{\mathrm{move.role}}_{s[i]}).$$

Figure 7.20 shows the estimates for the intonational predictors. We see significant positive effects for LC2 and LC3 (estimated coefficients: 0.12 and 0.2 respectively.) However, the magnitudes of these effects are relatively small compared to the effects of role and move. For example, when LC2 equals 1 we get an approximate 3% increase in probability of a switch, where the 95% interval of values in the observed data for LC2 is $(-4.15, 2.30)$ (LC3: $(-1.39, 2.12)$).

So it seems that having higher tilt or convexity nudges up the probability of a speaker switch, but the contribution is not as strong as that of the contextual cues.

[14] 'Divide by 4 rule': x on the log scale is at most $x/4$ on the probability scale.

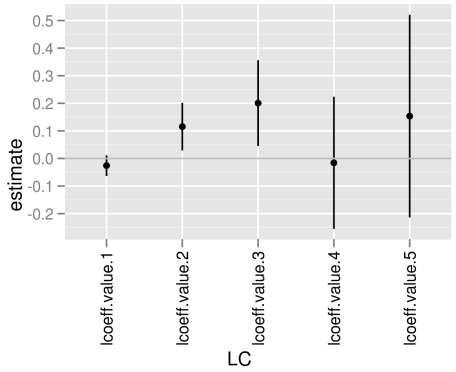


Figure 7.20: With LC data as individual level predictors.

In terms of role, this probably reflects the fact that the most efficient completion of map task involves minimal content from the instruction follower. When the follower has to provide content it usually means the pair have encountered a problem. As we would expect, question type moves are generally turn giving irrespective of whether the utterance has a rising or falling boundary. This suggests that intonation may mean more on some move types than others in terms of turn-taking. To check this we add the LC features to the model of the move group level predictor, allowing the coefficients for the LC values to vary by move type. The difference between the previous two models is very small. DIC values for the model in (4) and the new model are 1809 and 1802, respectively. However, the new model adds many more parameters. Figure 7.21 shows the estimates for the move predictors when allowing varying slopes. We see that the effects of coefficients associated of the contour features are very small compared to move type intercept. Out of the non-questions, only the affirmatives have a positive coefficient for LC2 and LC3. So, we don't see much of an interaction between move type and intonation in predicting speaker switches.

Finally, we might assume that the inference from rising intonation is that the speaker requires ratification of their move. For Inform and Instruct moves this would mean an Affirm move. That is, if rises signal a need for ratification we would expect to see that the probability of an affirmative increase with LC2 and LC3. To look at this, we fit a multilevel logistic model predicting whether or not Inform or Instruct moves are followed by Affirm moves. This is basically the same model as the previous speaker switching, but coding an Affirm response as 1, and other responses as 0. After controlling for the higher level dialogue features, we see that the effects of the contour shape features to be, again, dwarfed by the effect of role. Estimated coefficients for the shape features are around ± 0.05 for at the move level, resulting in about a 1% increase or decrease in the likelihood of an Affirmative response for every LC coefficient unit

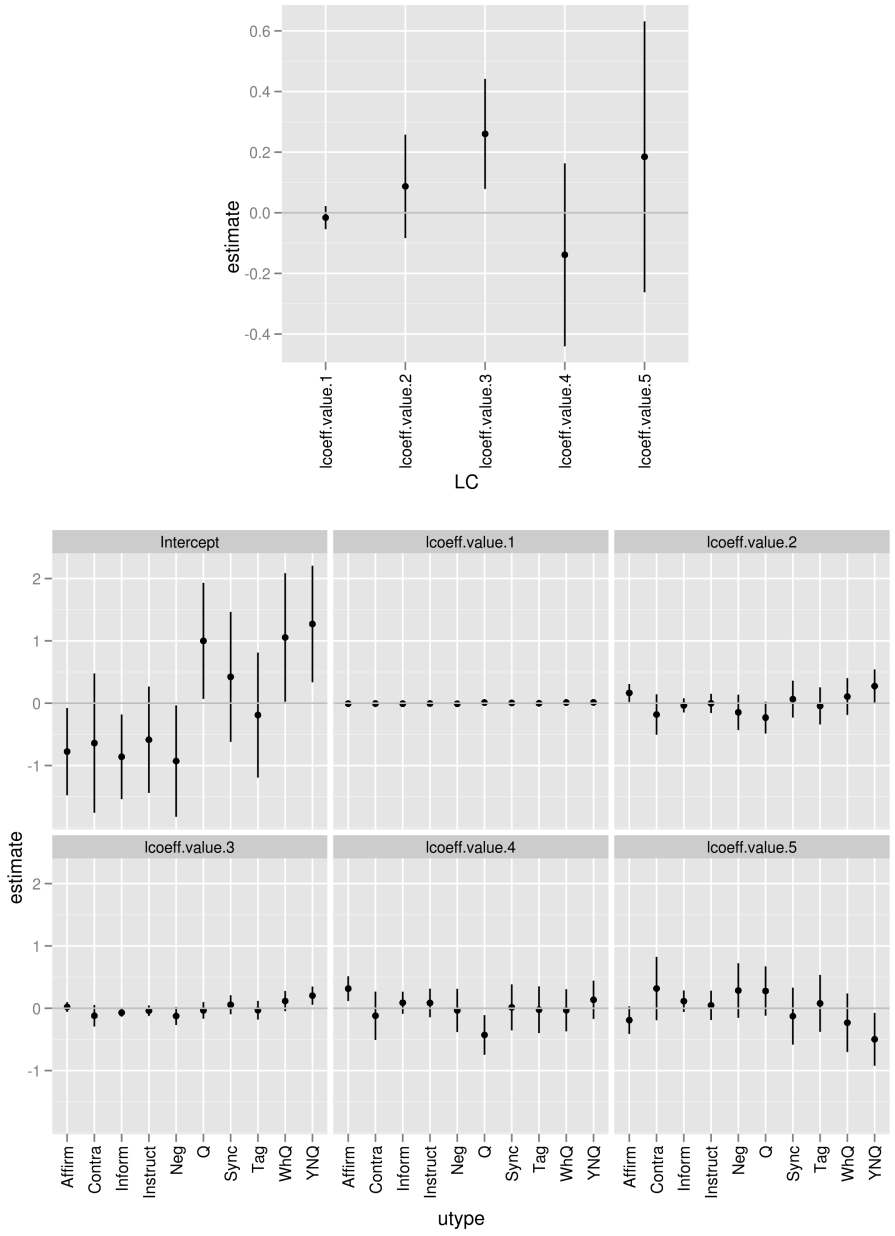


Figure 7.21: With LC data as individual level predictors and predictors on the move group.

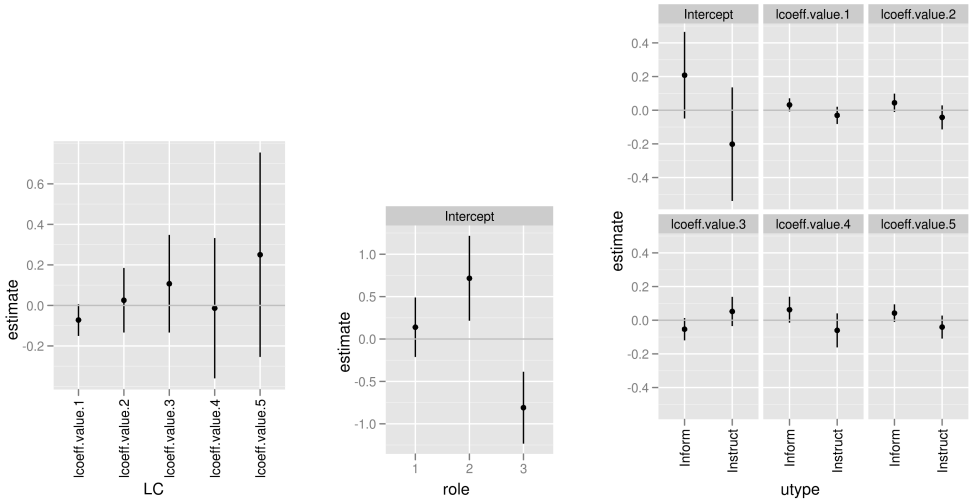


Figure 7.22: Predicting Affirmatives after Inform and Instruct moves with LC data as individual level predictors and predictors on the move group.

(cf. Figure 7.22). The effects are similarly close to zero at the individual level. On the other hand, being the map task follower, again, increases the probability of of the next move being an affirmative by about 18%. This indicates that, in fact, an Inform/Instruct by the *follower* is more likely to be explicitly affirmed than that of the giver. This is due to the fact that in the other two roles the speaker is likely to follow up with another content move. In general, we don't see that shape features on content moves are predictive of whether or not that move will be explicitly ratified.

Summary: Turn-taking

We previously said that if rises were turn holding we would expect to see more of them in the conversational dialogues. This is clearly not the case. In fact, we see more rising features in the map task data, which is also the data set with relatively more speaker switches. When we examined the relative effects of contextual factors, like role and move, and intonational shape features, we found that the former were more predictive of whether or not a speaker switch will take place, than the pitch shape features. We did find positive relationship between LC3 and LC2 and speaker changes. This would suggest that rises are mildly turn-giving which is congruent with there being more in the map-task. However, the effects are relatively small compared to role and move. So, we would not like to say that the reason that we see more rising features in the map task data is simply due to the fact that there are more switches. Similarly, it does not seem that rises are to be thought of as ratification or feedback

requests for Inform or Instruct games. So, it seems we need to look for reasons why there are more switches in this type of dialogue in the first place.

7.5.5 Discussion and Summary

Our goal was to find out if higher level effects like task/role had an effect of whether an utterance was produced with rising features. Comparing the feature distribution with that of read utterances, we see that Instruct moves are in the fall-rise space of the LC3-LC2 plane. So, it seems that we do get more rising features in the task-oriented speech than the free conversation. While we saw that significant positive effects of tilt and convexity coefficients for speaker changes, we also saw that these effects were dwarfed by the effect of role. So, whatever the rising features are doing on these turns, it does not seem that their main role is to manage turn-taking. Instead, it seems that the turn-taking strategy is mostly dictated by the higher level, task structure of the dialogue. The most efficient completion of map tasks would involve the least input from the follower, so the floor remains with the instruction giver as much as possible. However, to complete the task, the follower does have to keep the common ground synchronized with instruction giver. This is reflected by fact that affirmatives were much more common in the map-task than in the conversational speech.

How do our results fit with other analyses of rises? Once again it is hard to reconcile the data with accounts that link rises to *propositional* attitudes. For example, we wouldn't want to associate these rising features with how we usually think of epistemic uncertainty (contra Hirschberg's (2002a) Maxim of Pitch). If the instruction giver is uncertain about anything, it's not about the actual route. Instead it seems to be about whether the follower can or will carry out the task, i.e. discourse structural uncertainty.[15] Similarly, we don't want to associate rising features with submissiveness (Merin and Bartels, 1997) or lack of speaker commitment (Steedman, 2000), since these features are predominantly used by by the instruction giver, i.e. the situationally dominant speaker.

Still, the flavour of submissiveness that so often associates with rises (similarly uncertainty and hearer responsibility) can be made to fit with the findings when we recall that dialogue participants are involved co-operative tasks. For example, the follower needs to carry out the instruction for the map task to proceed. In this sense, the instruction giver is at the mercy of the follower. This sort of thinking brings us to the contingency or forward dependence type analyses of rises of Pierrehumbert

[15]My hunch is that slowing of speech rate and lengthening especially of consonantal regions should be a better indicator of actual uncertainty, because its more inline with other physical signals of tentativeness. We might test this by doing a joint task, by telling the subject not to perform the suggested action if the collaborator sounds too unsure.

and Hirschberg (1990), which is basically put into a more formal form in Gunlogson (2008). In this sort of analysis, rises signal that the move associated with a rise is contingent on whether or not some other discourse condition gains purchase. So, for example, Gunlogson argues that rising declaratives (i.e. with nuclear rising accents) signal that an addition to the common ground will only happen if the hearer ratifies it.

Gunlogson's work only really claims to deal with utterances with nuclear rises. However, it does seem like we could stretch this idea to terminal rises in general (as originally claimed by Pierrehumbert and Hirschberg (1990)) and by doing so capture the submissive/hearer responsibility analysis of rises. For current data, we might say that adding an action to someone else's to-do list requires acceptance from that person. So, at a glance we might say that the map-task shows more rising features because it just has more contingent elements. However, for the map task, there is no reason that we should take an instruction with a rise to be more dependent on affirmation than one without. In general, if we take instructions to be contingent we have to basically allow every move that adds content to be contingent, e.g. Inform/statement moves in the conversational speech. So, we still have the two problems of why some content moves get rises and some don't, and why there are more rising features on the map-task instructions. We cannot explain these differences by taking rise-contingency to mean that we need an explicit ratification. If it did, we would have seen a stronger connection between rise features and affirmative responses.

Given the high rate of affirmatives in the map task, I would argue that the reason we get more rises, particularly for the imperatives, is because co-ordination is more important for this sort of dialogue. The speaker deploys rising features because it is important to attend to whether a task is open or closed, since each subtask is dependent on the subtask before it. That is, we get more rises because there is a more well defined subtask structure (cf. Swerts and Geluykens (1994)). In this way the instruction giver is holding the floor at a higher level than the utterance-to-utterance level usually consider in studies of turn-taking. Similarly, the contingency we want to get at is at a higher level than accepting single instructions. At that lower level, speaker changes are possible, but in the end participants have to get back to the task sequence the giver is providing. So, instructions get more rising features because the type of dialogue game they arise in is primed for them.

This highlights the difference between non-finality point view of rises we have been running with and 'forward-looking' analyses in the mode of Pierrehumbert and Hirschberg (1990) or Gunlogson (2008). The former doesn't claim that the evaluation or action associated with the rising utterance necessarily depends on an *utterance* to come. So, a mid-task instruction with a low rise isn't really more dependent on future moves than one with a fall. Instead, the sort of dependency we observe is one that

points back up into the task/QUD structure. In contrast, in the type of conversational speech recorded in IViE there is less incentive to stay on track and there is less need for actual co-ordination. Speakers have more flexibility in leaving questions open. Put another way, speakers don't have so much reason to keep pointing back to unfinished business.

In general, we can think about the difference in dialogue types not only in terms of the actual task structure but also how far a speaker's actual beliefs/model of the world is allowed to differ from that projected by the common ground. This is determined by how important the quality of the common ground is to achieving the goal of the discourse. An extreme example would be when the goal is simply to fill silence (i.e. small talk). In this case, the hearer-as-evaluator may not care to raise objections to propositions they disagree with as long as the someone is talking. On the other end of the spectrum we have situations where, for example, identifying the wrong object may be extremely costly to the parties involved. The distribution of rises seems to reflect these differences and so could be a useful predictor of the dialogue quality. However, a greater range of dialogue styles would be needed to test investigate this.

Given the limited annotation, we could not model all the factors we would like to. In particular, a more detailed analysis of the actual task/question-answer structure will have to wait for the future. We would expect the QAC related factors examined in Chapter 5 to have an effect on boundary intonation and nuclear accent shape. However, we should note that the dialogues in Chapter 5 are quite different to the spontaneous speech here. Those dialogues were constructed as more task oriented in the sense that the contexts were about people trying to come to common ground about an event that had happened in the past. So, again, we would expect to see some differences between the findings reported in that chapter and what we would see in the IViE style conversations. That is, the local discourse factors will be modulated by higher level task effects.

So, the take home point at this stage is that higher level discourse factors, like task and role, have an effect on whether an utterance is produced with rising features or not. The more speakers have to co-ordinate through verbal signals, the more rising features we expect to see. So, these sorts of contextual conditions need to be taken into account when we collect and analyze intonational data. While the data here came from a single dialect, we also saw individual speaker level effects. This leads us to the more general question of how factors outside the discourse structure affect the production of rising features. We look at this issue in the next section in terms of dialectal variation.

7.6 Dialectal Differences in Read Speech

So, far we have been looking at the relationship between elements of the discourse and utterance final rises. We now turn our attention to how dialect, as a speaker intrinsic factor, affects this sort of boundary intonation. To do this we will look at the collection of read utterances in the IViE corpus. In particular, we would like to see how stable the relationship between sentence/move types and boundary prosody is across dialects. This aspect of the corpus has already studied both in a series of works in terms of features derived from the signal (Grabe et al., 2003) and nuclear tone categorization (Grabe, 2004). They found that differences in the intonation of between dialects could be captured quantitatively using Legendre polynomial decomposition. Moreover, that analysis indicated that contours were very similarly shaped across utterance types for Belfast and Leeds speakers, but not for the other regions. However, average F0 was higher in declarative questions than in statements across the regions. For the tonal analysis they found that move types like 'question' exhibited several different nuclear tunes. Moreover, more types of nuclear tunes were observed in northern dialects than in the south.

So what is there new for us to do? It turns out there are several things. First, the previous works looked at Legendre coefficients derived over the entire utterance. While there is obviously something to be gained from looking at F0 trends on a larger scale, this can also obscure the details of what's happening at the boundary, particularly the convexity of the nuclear accent.[16] This can be problematic if we want to generalize to longer utterances. So, we would like to know if distinctions can be made based on this smaller target area.

Our second point of interest is in the distributions of intonational features. The analysis in (Grabe et al., 2003) presents the data in terms of medians, but we aren't given much of an indication of how variable the data was. However, the results of previous tone annotation often presents several tones for some sentence types. This variation is shown in Figures 7.23 and 7.24 which depict the proportions of nuclear tune annotations for the different regions and move types (Grabe, 2004). So, we would like to know how this variation relates to the distribution of Legendre coefficients. For example, are there wide differences within tone categories? Do the differences relate to differences in region or utterance type factors?

Finally, since our analysis is not based on tone labels, we include all the speakers recorded. This doubles the amount of data analyzed from the previous studies. Through the analysis of this data, I will argue that we can get to the same generalizations about dialectal variation that we would looking at categorical tone labels.

[16] Remember, the earlier in the series, the more global the interpretation of the coefficient.

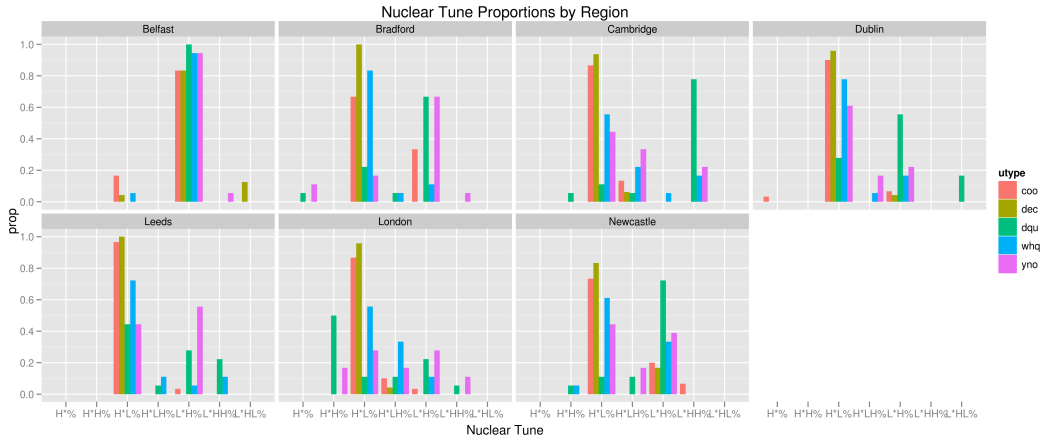


Figure 7.23: Nuclear Tune Proportions by region (cf. Grabe (2004)).

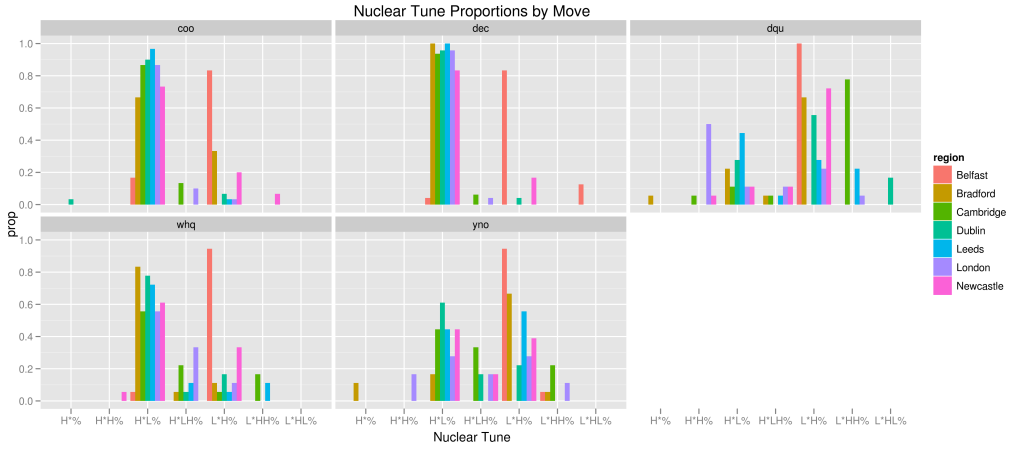


Figure 7.24: Nuclear Tune Proportions by move type from (cf. Grabe (2004))

In fact, we can give a better account of the relationship between region, utterance type and boundary intonation, by looking at the magnitude of the intonational gestures, Something that is not possible with ToBI style labels, but easy with continuous parameters.

7.6.1 Shape differences

Figure 7.26 shows mean contours based on Legendre polynomial coefficients over the utterance final word for each sentence type. The results are congruent with previous observations of Belfast utterances having a default boundary rise. All of the sentence types appear to have the same convex rising shape, although there are clearly differences in the size of the gesture. We can see that the declarative questions

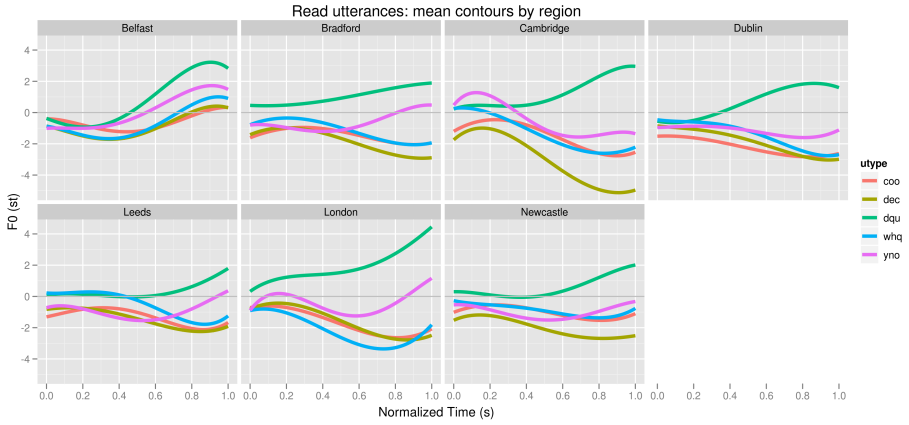


Figure 7.25: Mean Contours by region (last word)

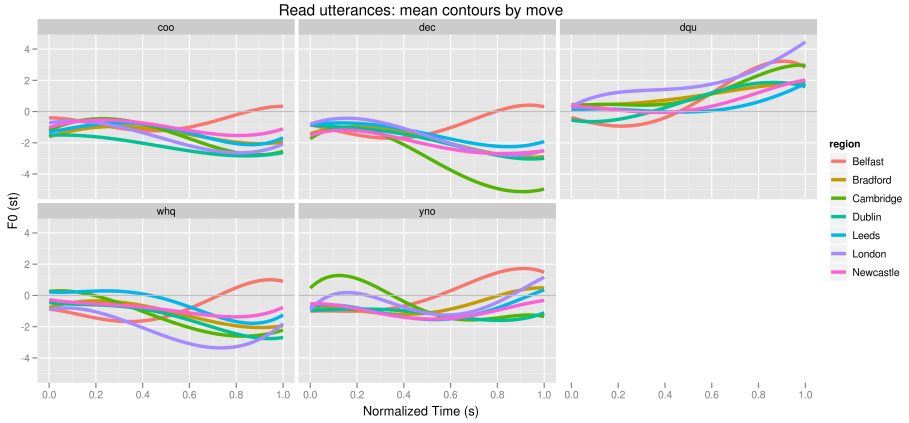


Figure 7.26: Mean Contours by sentence type (last word).

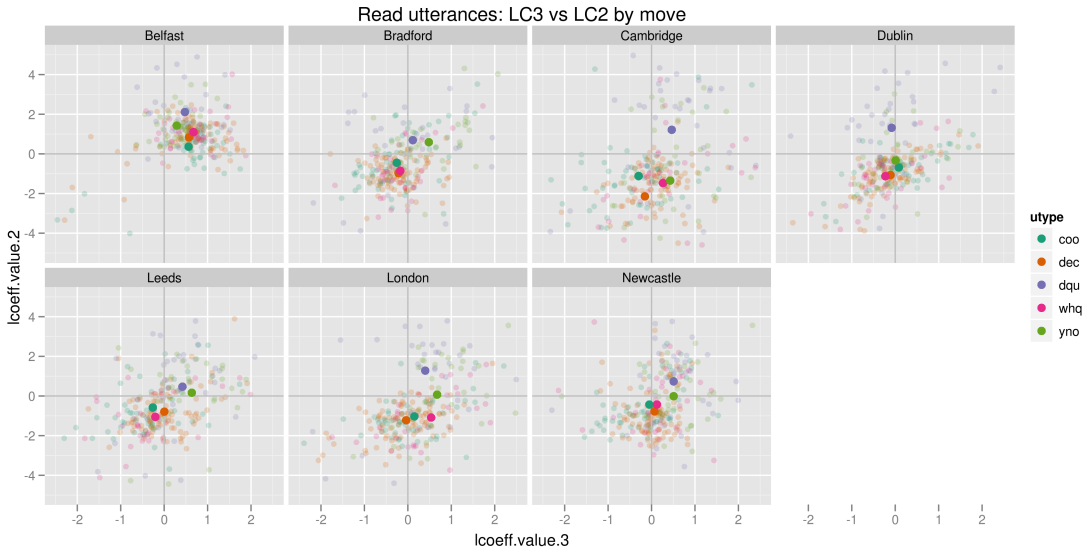


Figure 7.27: LC3 v LC2, by region (last word).

have, on average, a bigger rise than the other sentence types. The mean contours for declarative questions show a similar rising shape across regions, although the Dublin rise seems to be more of a plateau. Interestingly, we do not find support for the claim in Grabe et al. (2003) that Leeds contours do not differ much between sentence type. We see something like traditional distinction in between polar and declarative questions and the other sentence types, with contour for the former set seems to generally end higher. This extends more generally through the regions.

Figure 7.26 groups curves by sentence type. The declarative statement fall is most pronounced for the Cambridge data. This view also highlights the similarity between declarative statements that are not from Cambridge or Belfast. Declarative question is the category that exhibits the most cohesion across all regions. We can also note that the declarative questions from Belfast are distinctly higher than their statement counterparts. Similarly, the Cambridge polar questions are falling, but not as much as statements.

Exploring the space

It is instructive to look at the actual distributions of coefficients in more detail. In particular, we are interested in mapping out the relationship between move type and region with respect to convexity, which is not much discussed in the previous IViE works. Figures 7.27 and 7.28 show the data projected on the LC3-LC2 space with means. From these graphs, we can see how the distribution of the data in terms

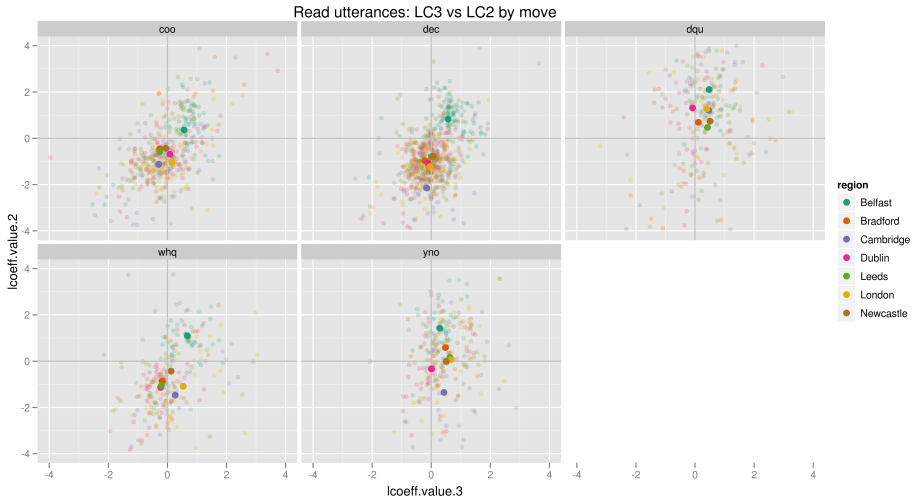


Figure 7.28: LC3 v LC2 with means, by sentence type (last word).

of function coefficients relate to the mean curves shown above. We see that the Belfast data inhabit the upper right quadrant of the plane, indicating positive tilt and convexity. The main difference for this dialect is in terms of tilt (similarly height, LC1).

Bradford and Leeds show more of the traditional question/statement distinction, with the means for statements, wh-questions and alternative questions in the lower left quadrant, while declarative questions and yes/no questions reside in the upper right. The other northern set from Newcastle also shows the same sort of grouping although the non-polar questions are closer to the origin. For the Cambridge data, the yes/no questions and wh-questions sit over in the lower right (+/-) quadrant, a little bit closer to fall-rise territory. So, we see the Cambridge data as falling at the boundary in more move types.

Last word versus whole contour

These observations differ from those presented in Grabe et al. (2003) who, for example, don't find Leeds declarative questions to be rising. So, how much difference does using the whole utterance contour make? Figure 7.29 shows the time normalized mean contours based on Legendre polynomial decomposition over the whole utterances. We can clearly see that things are happening before the utterance end which might help us distinguish the contours of the different sentence types. Most saliently, polar and wh-questions seem to have peaks on the auxiliary and wh-word respectively. However, estimating Legendre coefficients over these extended domains comes at the cost of how accurately we can model boundary prosody. Figure 7.30 shows the mean

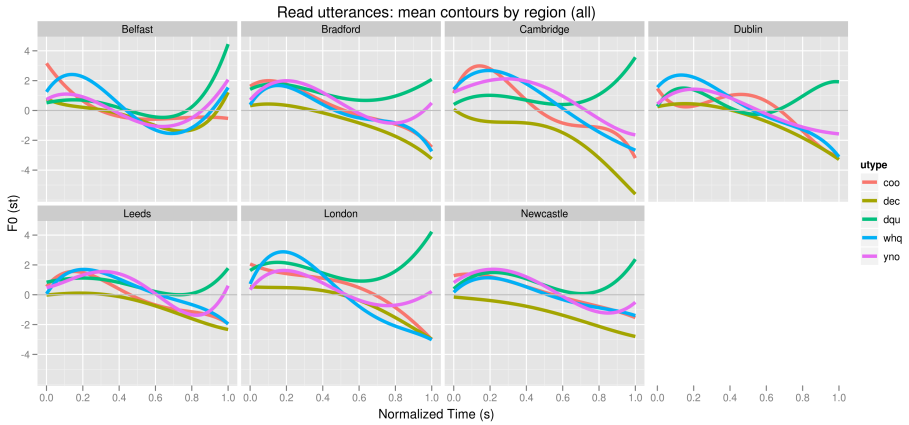


Figure 7.29: mean contours: whole utterance, by region

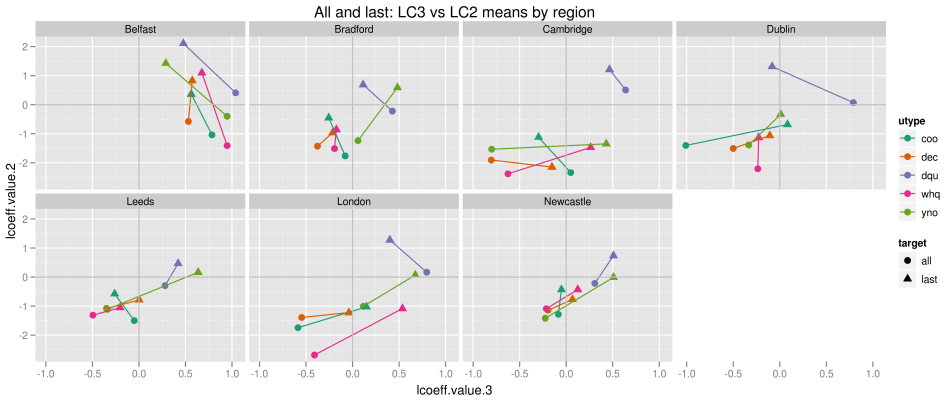


Figure 7.30: Last word and whole utterance means

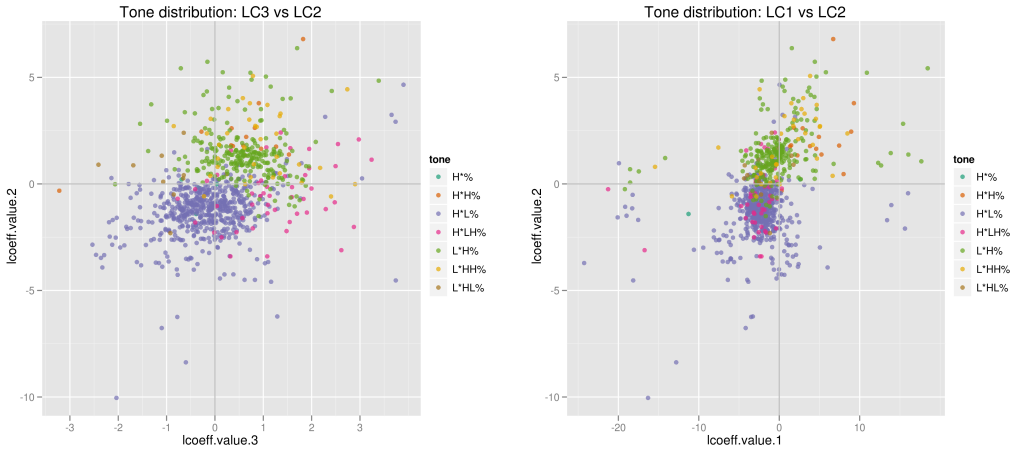


Figure 7.31: IViE Tone Labelling: LC3 (convexity) v LC2 (tilt) and LC1 (height) v LC2 (tilt).

LC3 and LC2 values based on these two target sizes. The main difference for the Belfast data is that tilt LC2 values go from being negative to positive. Similarly, the means for the polar questions increase in both dimensions across. So, focusing on the last word allows us to better see the difference between questions and statements at the boundary.[17]

In general, we see quite clearly see that boundary intonation is grammaticalized quite differently from between regions. For speakers from default rising dialects like Belfast, dialect/region should be a strong predictor of whether or not an utterance will have with these rise features, beyond the discourse level factors discussed in previous chapters.[18] The situation is more variable for the 'falling' regions, where we see differences in the sign of tilt and convexity matching differences in sentence/move type more clearly when we focus on the boundary. In the following sections we will look more closely at the connection between our derived features and the ToBI style tone annotation. I will argue that that such labels only really communicate the sign of shape parameters. However, for the issues we are interested in we really need to take into account their magnitude.

Figure 7.32: IViE Tone Labelling: LC3 v LC2. Notice the green hair on the L*H% cluster!

7.6.2 Connection to ToBI style labels

How do the Legendre polynomial coefficients relate to the manual annotation of nuclear tones? Figure 7.31 shows the projection of the tone labels onto the LC3 v LC2 and LC1 v LC2 planes.[19] The LC3 v LC2 projects maps quite well to our intuitive expectations. The upper right quadrant is the home of the rises (L*H%, L*HH%, HH%). Fall-rises (H*LH%) have relatively high LC3 values. Generic falls (H*L%) are generally negative in LC2, but spread out on both sides from the LC3 origin. The LC1 v LC2 map gives an indication of the distinction between L*H% (rise plateau) and L*HH% (rise proper), but we lose the distinction between fall (H*L%) and fall-rise (H*LH%). Figure 7.32 shows the clusters for the actual tones and move types, pooling the regions. Clusters appear where we would expect them, although there is clearly a lot of overlap between classes.

According to the original tonal results (Figure 7.23) sentence types manifested

[17]You could also use both sets of features as predictors, and that's indeed what people have done for DA taggers using means and slopes.

[18]While we don't have time for a full investigation of dialogue in Belfast speech, it impressionistically it does seem to be the case that the rising carries over to the dialogues, but there is speaker variability in the monologues.

[19]Remember only half the data was annotated.

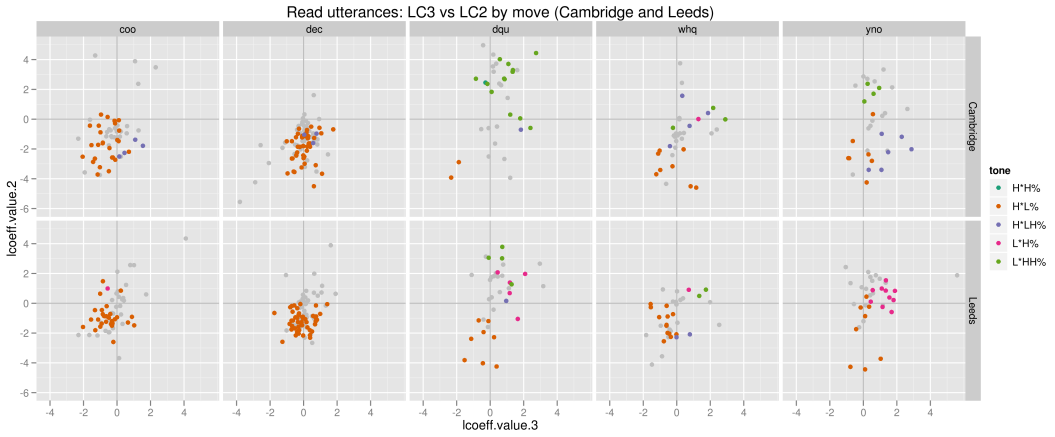


Figure 7.33: LC3 v LC2 by sentence type and tone: Cambridge and Leeds. The grey points represent the unannotated data. Some outliers were removed for better viewing.

with sometimes very different tones. For example, for the Cambridge yes/no questions 44.4% were falling (H*L), while 27.8% had a fall-rise tone (H*LH%) and 27.8% had a rise proper (L*HH%). Similarly the Leeds yes/no questions were split 44.4% to 55.6% between falling (H*L) and rising (L*H%). Figure 7.33 shows these differences recast in Legendre coefficients. The clusters for the different tones are indeed where we would expect them to be for the yes/no questions. Adding the unannotated data (the grey points), we see that the majority of Leeds yes/no questions reside in the positive L2 space, i.e. they are more rising.

Looking at the distributions, we can also see why looking at medians might lead Grabe et al. (2003) to suggest that declarative questions were not rising in Leeds: the annotated data are more or less evenly distributed in the upper right and lower left quadrants (i.e. rising and falling), leaving the median in close to zero and similarly close to the median for the declarative statements. We see the rise in our mean curves (Figure 7.29) because we include the unannotated data, the majority of which reside in the upper right quadrant. Note also, that the variance for the declarative questions is much larger than that for the statements. We take this as a cautionary tale about point estimates.

The borders between tone groups are not cleanly delineated. In particular, we see three H*LH% labelled utterances, amidst the mass of H*L% Cambridge declaratives. Closer inspection suggests that those points really do belong there, based on their shape. Figure 7.34 shows the contours of three Cambridge utterances over the same final word ('Ealing'). Two of these are declaratives with very close LC2 and LC3 values labelled H*LH% and H*L% (c-dec1-f2: (LC2=-1.04, LC3=0.03), c-dec1-m3:

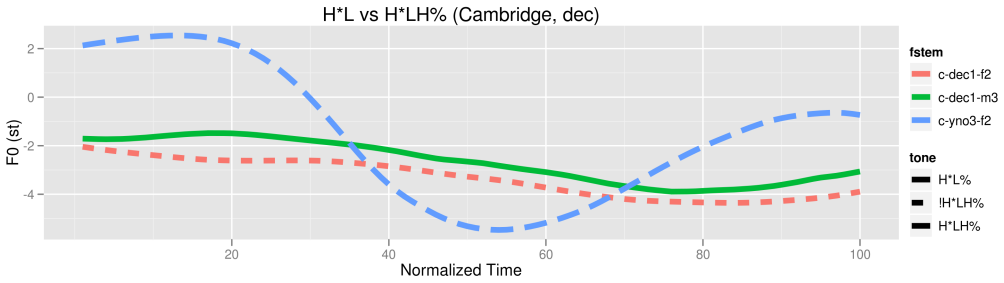


Figure 7.34: H*L versus H*LH% in two Cambridge declaratives and one polar question. The target word in each case was 'Ealing'. Note the tone transcription includes downstepped tones, but this distinction is not kept in the results of Grabe (2004).

(-1.14, 0.06), respectively). The other is a polar question labelled H*LH% from the declarative H*LH% speaker (c-yno3-f3: (-2.02, 2.87)). The two declaratives are clearly much more similar than the polar question, even though the tone labels would suggest otherwise. On close auditory inspection the H*LH% declarative does sound like it has a very small rise on the end, while H*L% declarative doesn't. So, the small uptick in the latter is likely slightly misleading. However, in my opinion, both declaratives sound predominantly falling, whereas the H*LH% polar question is very saliently a fall-rise. This seems better captured by the distance between their features, rather than the tone labels.[20]

In sum, we can come to the same sorts of conclusions about the distribution of boundary intonation/nuclear tunes using the automatically derived contour features as for the manual annotation of labels. Moreover, using continuous rather than discrete variables gives us a better idea of how close contours are to one another. This, of course, depends on us knowing how to interpret the feature space. Clearly, having some tiny positive value of LC2 won't necessarily map to the perception of a rise. However, it does seem clear that as LC2 and LC3 values pass a certain threshold, perception of a rise becomes inevitable. To better map out how the functional feature space relates to perception of contour shape we need more data, in both production and perception terms. We leave this as future work, but note that it seems more likely that we will get a fuller mapping using controlled contextual variations rather than isolated utterances, especially when it comes to fall-rise type contours (cf. Chapter 5).

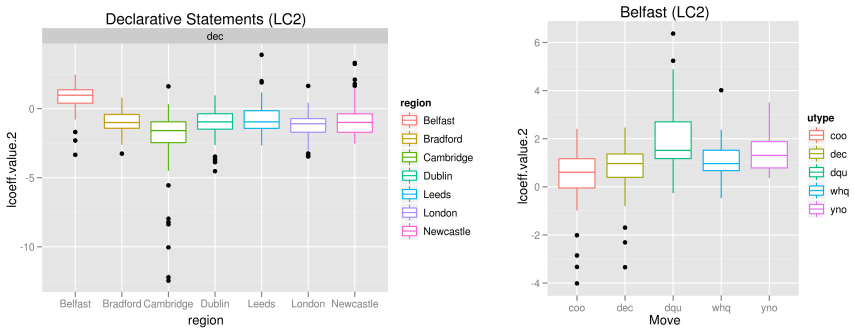


Figure 7.35: Declarative Statements (left), Belfast data (right)

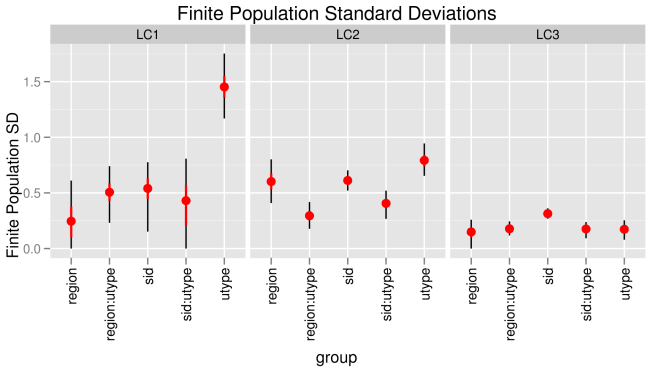


Figure 7.36: Finite population standard deviation estimates (s): medians, 95% (black) and 50% intervals (red) from 1000 MCMC samples

7.6.3 Beyond Labels

As noted previously, Cambridge declarative statements appear to have steeper utterance final falls than those from the Northern regions, even though around 94% of all such tunes are have the same H*L% label. Similarly, declarative questions in Belfast English seem to rise higher than declarative questions, even though the vast majority of these utterances were labelled L*H% (cf. Figure 7.35). So, it seems that magnitude of the gesture rather than orientation distinguishes region in the former case, and move type in the second. These distinctions are not captured with ToBI like tone inventories. However, they easily come out of a more gradient analysis.

We model the effects of speaker, move, and region on the first three Legendre coefficients as follows:

$$y_i = \alpha^{\text{region}}_{j[i]} + \alpha^{\text{spk}}_{k[i]} + \alpha^{\text{move}}_{l[i]} + \alpha^{\text{move.region}}_{s[i]} + \alpha^{\text{move.spk}}_{t[i]}. \quad (5)$$

[20] Clearly some of the derived feature values are artefacts due to errors in pitch tracking. The point is that with enough data we can admit some tolerance for such errors.

The group level coefficients were modelled as follows:

$$\begin{aligned}
\alpha_j^{\text{region}} &\sim N(0, \sigma^2_{\text{region}}) \text{ for } j = 1, \ldots, 7 \\
\alpha_k^{\text{spk}} &\sim N(0, \sigma^2_{\text{spk}}) \text{ for } k = 1, \ldots, 84 \\
\alpha_l^{\text{move}} &\sim N(0, \sigma^2_{\text{move}}) \text{ for } l = 1, \ldots, 5 \\
\alpha_s^{\text{move.region}} &\sim N(0, \sigma^2_{\text{move.region}}) \text{ for } s = 1, \ldots, 5 \times 7 \\
\alpha_t^{\text{move.spk}} &\sim N(0, \sigma^2_{\text{move.spk}}) \text{ for } t = 1, \ldots, 5 \times 84.
\end{aligned}$$

Finite population standard deviation estimates are shown in Figure 7.36 for each of the models predicting the first three Legendre coefficients. These values were derived from 1000 samples from the posterior distribution of the parameters of the fitted model using MCMC methods (R coda: mcmcsamp) We see that the variation between estimates for the different move types is relatively large in the LC1 model – much bigger than between regions. The regional variation is greater looking at the tilt (LC2), but still probably less than or equal to that between move types. We see less variation between predictors within our groups for convexity (LC3). This suggests that variation on this dimension is would be better modelled with other features. Given the findings in the previous chapters, it seems reasonable to suggest that these other factors may be more related to the task/role structure, or speaker engagement. However, in the context of these isolated read utterances, we see differences mostly relayed by height.

Figures 7.37, 7.38 and 7.39 show parameter estimates for the actual predictors. As we would expect, we see a large positive effect for declarative questions and a negative effect for declarative statements. Looking at the interaction between region and move for LC1, we see a (just) significant negative effect for Cambridge declaratives. The reason we see bigger effects for the bias term than for the tilt can probably be attributed to plateau effects for both rises and falls.[21] Looking at the relationship between region and LC2, we see a distinct positive effect for Belfast speakers. Again, the effects in the LC3 model are smaller. However, we still pick up that Belfast utterances have greater convexity in general, as do polar questions across regions. So, the fact that Belfast declarative question rises look bigger than declarative statements rises reflects a cross-dialect difference between these two move types. Similarly, regions vary in how low their falls are. In summary, we clearly see differences between and within dialects that are captured by a magnitude difference (i.e. pitch height, tilt) rather than differences in the sign of the tonal target (e.g. H*L% versus L*H%): we get more from looking at the at the data as continuously rather than categorically varying.

[21] Of course, the coefficients are correlated!

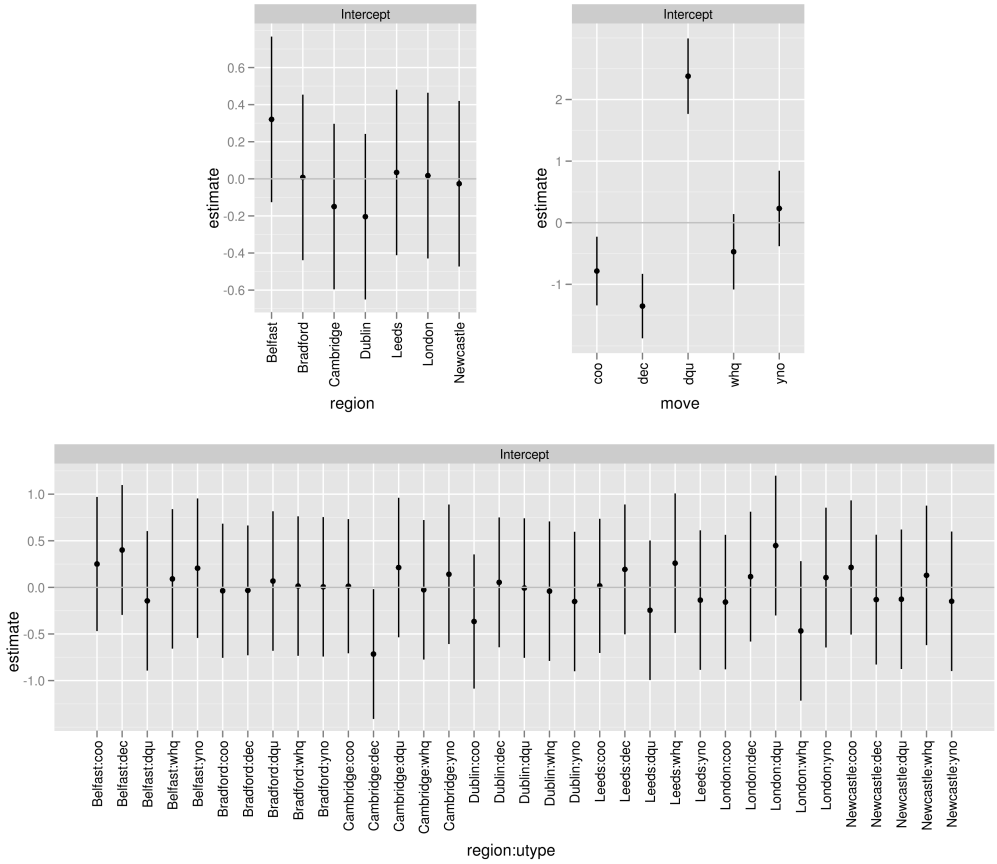


Figure 7.37: F0 height: LC1 model

7.6.4 Speaker Differences

Inter-speaker differences are also telling. Newcastle has been classified as a final rising region in several works (cf. Cruttenden (1997)). We don't really see this in the pooled statistics for the read utterances. However, Figure 7.40 shows some marked differences in the LC3 v LC2 distributions between speakers. In particular, we see speaker nm1 pretty much exclusively uses Belfast style final rises, while nm4 uses falling boundary intonation. In between these two, speaker nm5 seems to vary between having rising and falling statement declaratives.

This variation suggests that some speakers have grammaticalized the final rise as local boundary marking, while some have not. However, inspection of the dialogue data for these speakers suggest what is really happening is that some speakers vary more across speech contexts/styles than others. Impressionistically speaker nm2, for example, does exhibit some rise plateaux at the end of Inform moves in conversation (Figure 7.41). A more extreme example of this individual style shifting can be found

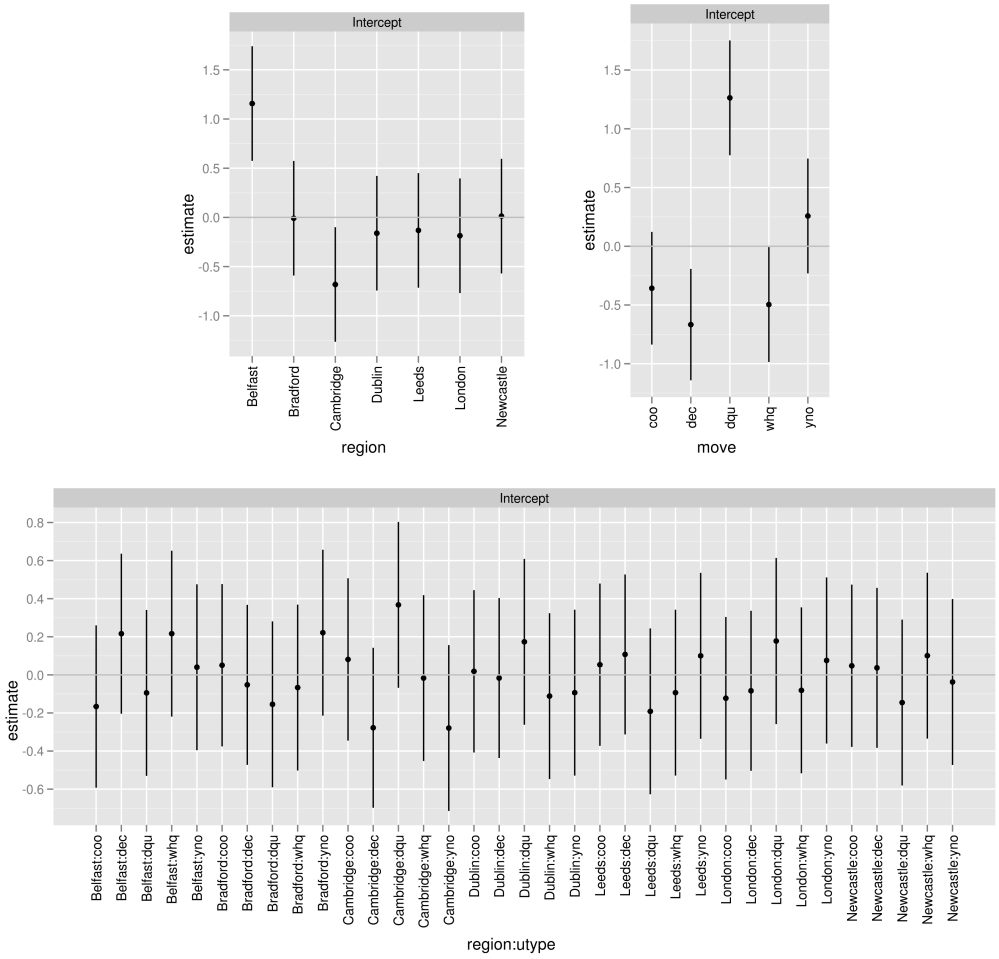


Figure 7.38: F0 tilt: LC2 model

in the Belfast data. Figure 7.42 shows that speaker bf3 consistently produced falling declarative statements. However, she seemed to consistently produce final rises in the conversational data (e.g. Figure 7.43).

It seems that the 'read' context speech is a predictor for some speakers but not for others. As far as I can tell, in the cases where there is some dichotomy for a speaker between read and spontaneous speech, then the read monologic speech will be biased towards falling. This fits with our previous argument that rises are indicative of inter-speaker co-operation. The upshot of this for our analysis is that when we see rises in isolated read declaratives (in English) we can assume that such rises are really grammaticalized as phrase delimiters, rather than signals of discourse attitude, in that dialect. Further investigation of the spontaneous dialogue portions of is clearly necessary to substantiate these. However, we leave this for future work.

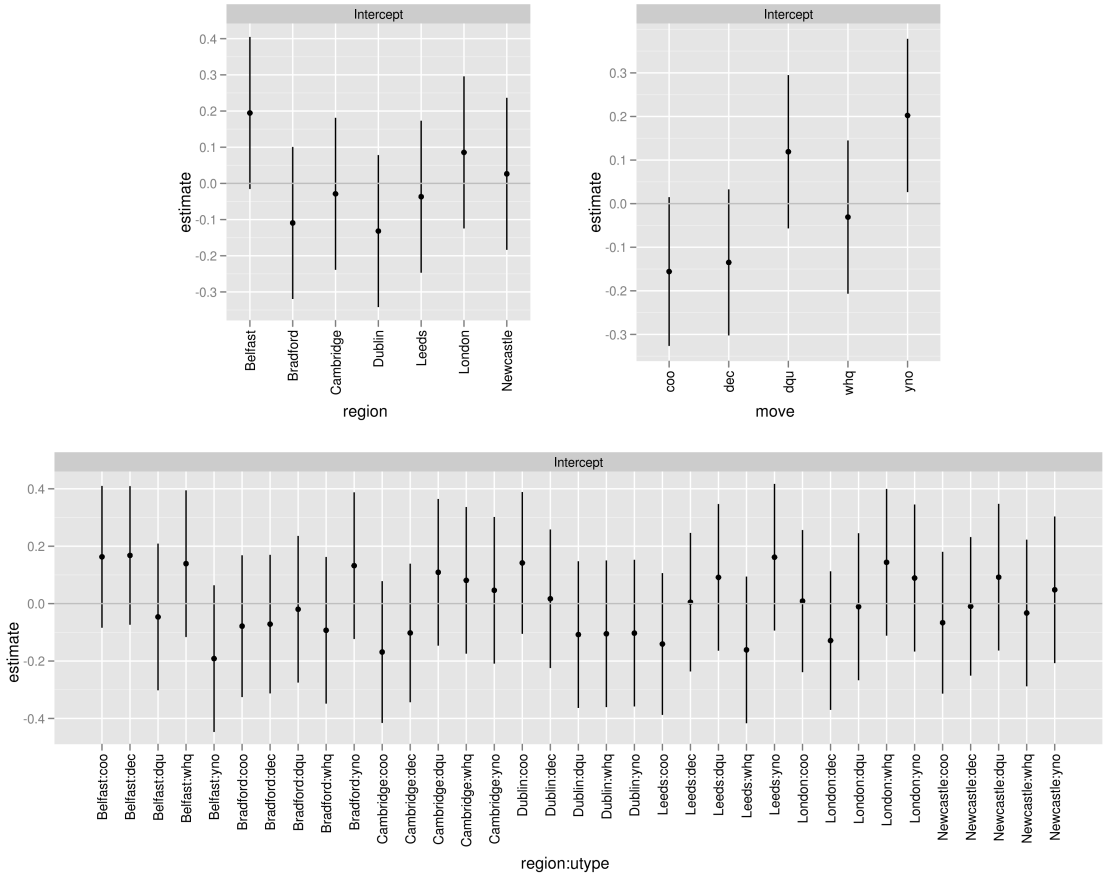


Figure 7.39: LC3 model

7.6.5 Discussion and Summary

We confirmed previous results from the IViE data, showing that boundary intonation is grammaticalized differently between dialects. The fact that the convex nuclear tunes are the default in Belfast means these sorts rises perform a different role than they do in Standard Southern British English. That is, we need to include these speaker level predictors in order to get the proper contribution of prosody. Nevertheless, the data suggests that we can still make cross-regional generalizations with respect to the mapping between intonation and move type. Most strikingly, we saw that declarative questions had higher end pitch than the other move types, across the regions. This observation had already been made in Grabe et al. (2003). However, that study analyzed shape parameters derived over whole utterances. This meant that some trends associated with the boundary intonation are not as visible. In particular, we found that polar questions were more like declarative questions, having higher tilt and convexity. This allows us to reconcile the results of Grabe et al. (2003)

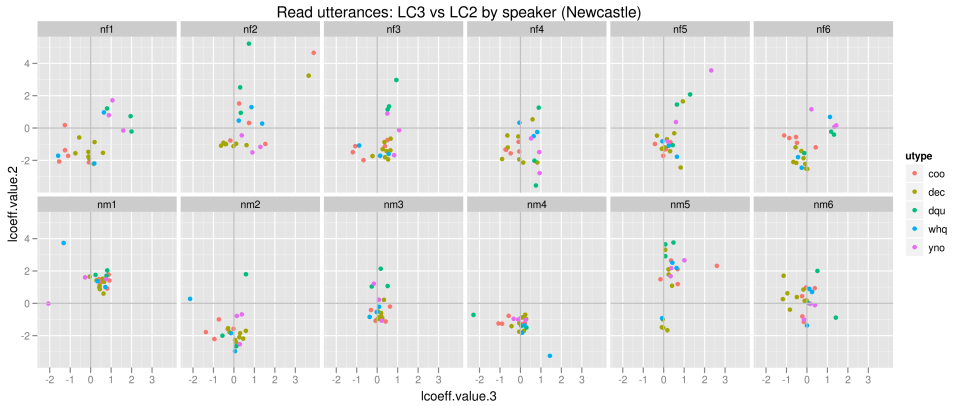


Figure 7.40: LC3 v LC2 by speaker (Newcastle).

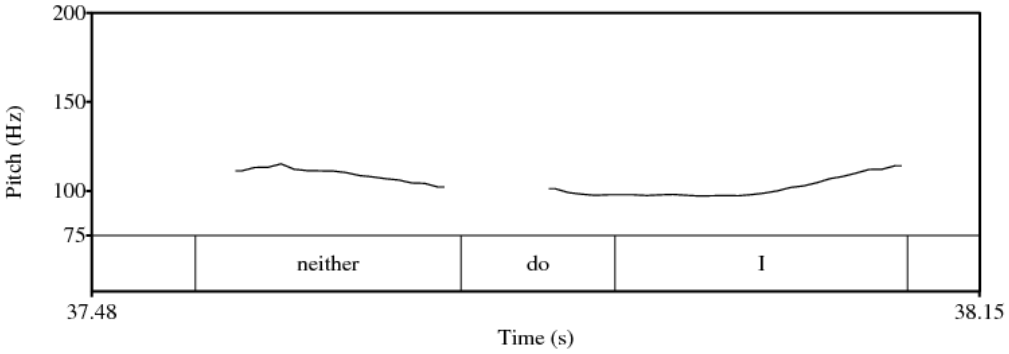


Figure 7.41: Newcastle nm2 in dialogue.

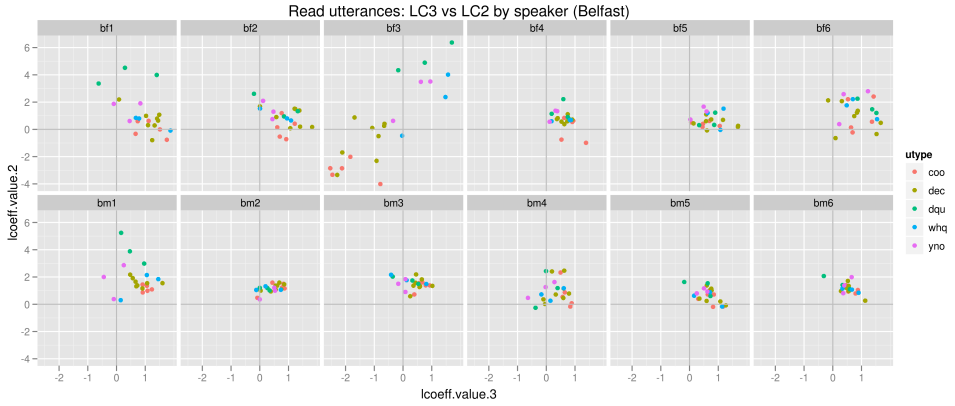


Figure 7.42: LC3 v LC2 by speaker (Belfast).

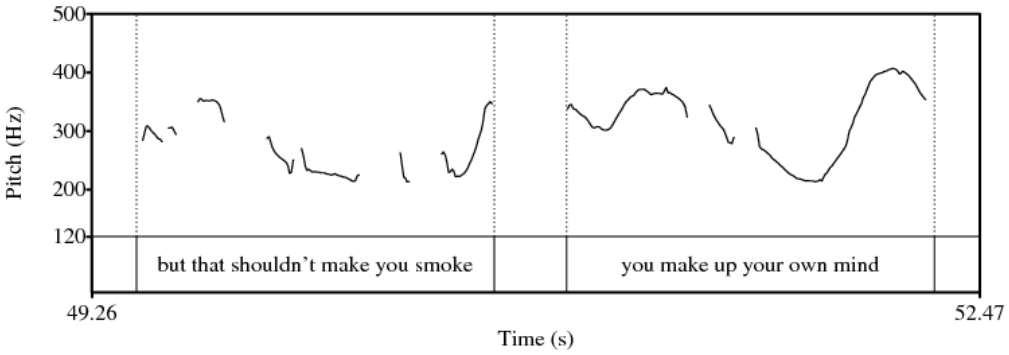


Figure 7.43: bf3 in dialogue.

with the results of categorical analysis of nuclear tones in Grabe (2004).

This work also treads on similar territory to Grabe et al. (2007) However, that work investigated Legendre coefficients as predictors of tone labels, while we looked at speaker level features as predictors of those coefficients. The former work argues that the categorical labels could be effectively replaced by continuously valued parameters. We take it a step further and show that the distinction between moves was better captured in terms of magnitude (e.g. height, slope) rather than abstract shape (i.e. *sign* of tilt and convexity). Beyond this, using these sorts of values derived from the signal gives us a notion of distance between contours which may be more useful in matching production to perception. The distribution of coefficients gives us a better idea of where the border territories really are. We would like to know, for example, whether having an upticking end on a contour with clearly negative tilt really counts as rising (cf. Section 7.6.2). Our data would suggest not.

Looking at the distributions highlights some important differences that are obscured when we just look at means or medians. For example, Grabe et al. (2003) reported that the intonational pattern of Leeds speech was quite uniform across sentence types based on median contours. However, examining the tilt and convexity coefficients (Figure 7.33) we see some marked differences between move types. In particular, there is a salient difference in the distribution of polar and wh-questions, with the former mostly located in opposite quadrants of the LC3-LC2 plane. More generally, we see much greater variance in the question categories than the statement category. While declarative questions mostly have a rising nuclear tune, we also observe distinct falls and fall-rises within regions. This is inline with the idea that when the meaning of the utterance already carries a strong signal of non-finality, e.g. interrogativity, more variability in the boundary intonation is possible. We previously saw this for the cue word *really* (Chapter 4) and for indirect declarative responses (Chapter 6).

Of course, this doesn't mean that we can now just pack all the variation into some random error term. Given our results on the indirect statement responses, it seems very likely that these differences reflect other speaker situational variables like speaker engagement and, of course, discourse structure. In fact, these isolated productions really only reflect one type of discourse situation: one where the task is atomic, non-interactive and doesn't require co-operation. Since the question under discussion is not explicitly defined, speakers probably default to broad focus, but we can't know for sure. In this sense, the discourse context is only very loosely defined. However, we have already seen in Chapter 5 that local contextual constraints (i.e. question-answer congruence) can have a big effect on the intonation of declaratives. The relative lack of fall-rise type accents in this data set is probably an offshoot of this. Similarly, the declarative questions we collected from American English speakers in Chapter 5 were much more uniform across speakers. It is plausible that this is due to context setting. Clearly, it would be advantageous to further explore the relationship between move types, intonation and our feature set through more context controlled elicitation.

Despite these limitations, the read data provides a useful lower bound of the dialect effect on rises. It also gives us an inkling of how the different factors, discourse level and dialect level, are weighted. We get an idea of this considering the gap between our results and the many qualitative reports of default rises in the Urban British North, particularly in Tyneside English (Local et al., 1986). Listening to the conversational data, it seems that the reading task seems to put a dampener on production of rising features, which do in fact seem quite pervasive in the dialogues (cf. Figure 7.41). However, this rise expectation is more or less wiped out by the change in task (conversation versus reading). The default nature of rises in Belfast English appears to be much stronger. So, strong in fact that it ubiquitous in the isolated read utterances for all but one speaker. This leads us to conclude that for Belfast English, the boundary rise (the sign of LC2) reflects phonological, phrase local, structure, rather than the discourse structure. That is, we expect phonological expectations to weigh heavily for Belfast speakers. On the other hand, we expect that speakers from a reading-falling region like Newcastle would be much more under the sway of discourse/situational factors when it comes to the distribution of rising features. We would expect the results to be similar to those found for the Cambridge dialogues.[22] To confirm these hypotheses we need to a proper quantitative analysis of the dialogue data for these regions, but at this point we need to leave this to future work.

It's worth noting at this point that differences in how intonation is grammatical-

[22]In this case, we might expect big falls, like those we see for Cambridge speakers' declaratives, to convey some extra meaning component for the Northern dialects, since the expectation is for shallower, possibly upticking falls. Again, we don't have the data to test this at this time.

ized do not seem to pose too great a hurdle to interpretation. Hearers seem to be able to accommodate prosodic differences quite easily. That is, intonation is interpreted relative to expectations (cf. McLemore (1991)). Once you've identified a speaker as having a default rise, you no longer look to the general shape at the boundary intonation for discourse attitude information.[23] While the inventory of boundary shapes isn't so informative for some dialects, the relative size or effort associated with a gesture does seem to be informative. Put another way, a really high rise is unexpected, even for a speaker from a rising dialect. It seems that speakers intonationally mark things that they want checked, e.g. the content of a declarative question, but to get the 'marked' interpretation you need to have enough information about the speaker to figure out when something is unexpected. This depends on discourse structural and phonological factors to different degrees depending on the discourse situation.

7.7 Conclusion

In this chapter, we looked at the distribution of rising features in conversational and task-oriented dialogues, and read speech. For the first part we specifically looked at productions from the Cambridge subset of the IViE corpus. Previous studies of speech from this region suggest that content providing moves, like declarative statements, are a canonically falling (Grabe, 2004, Cruttenden, 1997). That is, the prior probability of seeing rises is lower, so this dialect would give conservative estimates of frequent rising features in dialogues are. Overall, we found that content providing utterances in map-task dialogues had greater convexity than those from the conversational dialogue. Most of this seemed to come from instruction moves which often had a distinct fall-rise shape. Affirmatives were also more rising in the map task although here we saw more of the rise-plateau variety. While the rate of Affirmative moves was concurrently higher in the map task, we didn't find any strong link between rising features – higher LC2 and LC3 – and affirmative responses, or more generally speaker switches or stays. I argued that this state of affairs sits best with discourse structural analyses of rises, rather than more notions like submissiveness or uncertainty. It seems that differences in the need for participants to co-ordinate fit the differences in the distribution of rising features.

We then took dialect as case study of the effect of non-discourse factors on the distribution of rises. Our results basically extend previous work describing the regional variation in Britain: Belfast English is a strongly rising dialect, while the other dialects under consideration appear to be default falling. However, variation was seen within the falling dialects, particular between the exemplar of Southern

[23] A possible experiment might look at how long it takes to accommodate these dialectal features.

British English, Cambridge, and the data from the northern cities, with the former showing bigger falls on declarative statements. We confirm a magnitude difference, height, rather than categorical shape, to be indicative of differences between move types across the regions examined. In general, greater variance was observed in the contours for questioning type moves. Examination of this data highlights the importance of examining the distributional differences between and within regional groups. In particular, we saw how looking only at point estimates, like medians, can mask the amount of actual variation in a data set (e.g. the Leeds data). In terms of visualization, looking at the tilt versus convexity plane (LC3 v LC2) fits better with intuitive notions of nuclear tune differences than height versus tilt.

In summary, the studies presented in this chapter have shown that we need to take these higher level factors into account when proposing or evaluating theories based on intonational data. Further study is required to fill out the map between the data projection onto continuous measures like tilt and convexity and human perception of intonation. This will hopefully allow us to better tease out the phonological and discourse factors that all travel on the intonational bandwagon.

Chapter 8

Conclusion

8.1 Summary of Findings

This dissertation has investigated the interpretation and distribution of rises with respect to utterance semantics, local discourse structural constraints, and higher level contextual features. The effect of utterance semantics was examined first through studies of cue word responses. I particularly focused on *really* as a cue word that exhibits an ambiguity between backchannel and question type categories in dialogue act annotated data sets such as the Switchboard corpus. The connection between rises and questions is central to many previous accounts of rise meaning. So, looking at this sort of utterance allowed us to look at the connection between rises and dialogue moves in real conversations while evaluating those theories.

The corpus studies and perception experiments presented in Chapters 3 and 4 indicated that contour shape, particular the rise characteristic, did not have much to do with whether a *really* was labelled as a backchannel or a question. Instead, we found that having more effortful prosodic features, e.g. higher pitch level, or discourse active features, e.g. eliciting an evaluative response, better explained when *really* was interpreted as a question. Further perception experiments also looking at affirmative cue words showed that properties that have been associated with rises more reflect the properties of the underlying content and the discourse situation. However, rises appear to generally signal discourse non-finality. This is best interpreted as a form of discourse instability: there is some dominant and salient question under discussion (or task) which has not yet been resolved. However, in the right conditions is can be interpreted as propositional uncertainty. Combining the findings of the current study and previous investigations of affirmative cue words, the account of rise meaning that best fits the data is one that says something about the discourse configuration rather

than the propositional content of the rise carrier.

This view of rises was upheld in studies of declarative responses. The production and perception experiments presented in Chapters 5 and 6 also showed the value of analyzing rise meaning in terms of a QUD driven dialogue model that incorporates notions of information packaging in the sense of Vallduví (1990). We saw systematic prosodic differences arise when considering response dimensions like whether an utterance directly or indirectly addresses the question under discussion. In particular, indirect responses can (and often do) elicit a terminal rise as part of a fall-rise accent in the IS ground. This fits with the fact that such responses, by virtue of not being direct, signal that more needs to be said. The parallelism between the rise meaning and the discourse configuration was brought out in a perception experiment, where we saw that having a fall-rise didn't really add to the perception non-finality, or other uncertainty related attitudes, beyond what we get from the response type.

We can understand why indirect responses get the prosodic form they do by noting they are the type of discourse configuration where a evoking of a strategy, in the sense of Büring (2003), is necessary to maintain discourse coherence. However, while adding the rise seems necessary to evoke a strategy in some cases, e.g. when the response has the form of a direct one, it's not necessary in others. So, we saw that rise interpretation depends on discourse expectations. It seems that the dominant factor determining these expectation appears to be question answer congruence. This is reflected in the information structure of the response which is realized primarily in metrical structure terms, i.e. pitch accent placement rather than shape. In cases where rise meaning parallels what the proffered content tells us about the state of the discourse, we get more variation in production, less variation in perception.

Chapter 7 showed how higher level contextual factors affect the distribution of rises. Examination of task-oriented and conversational speech from the (Cambridge) IViE corpus, showed more rises on proffering moves and affirmative responses in the task oriented speech. For the former, most of these came in the form of fall-rises on instruction moves. We can see this as a reflection of the overall task structure and objectives of the Map Task method used to elicit the task-oriented dialogues. In general, we expect to see more (non-question) rises in situations where the need to co-ordinate, i.e. maintain a quality common ground, is greater. So, rises can be taken as an indicator of how interactive and co-operation dependent a speech style is.

Examining read speech from the IViE corpus further, we saw that pitch magnitude differences, e.g. height, were more indicative of differences between move types across dialects than sign, e.g. fall or rise. In general, greater variance was observed in the contours for questioning type moves, which is again a case where rise meaning would parallel textual meaning. Dialectal studies also indicate how intonational form can reflect expectations at different levels of the grammar. In Belfast English nuclear

rises are default across move types even in isolated read utterances. Thus, we expect phonological expectations to weigh heavily for Belfast speakers. Newcastle speakers, on the other hand, generally showed a falling statement pattern for read speech, but a rising pattern in spontaneous dialogues. So, we would expect that speakers this region to be more affected by discourse/situational factors. In both cases, however, we see that very big rises are still marked/unexpected and hence lead to a check/question interpretation. This once again points to the importance of taking discourse and grammatical expectations into account when dealing with prosodic interpretation.

8.2 Where does intonation act?

Based on these studies, I have argued that the aspect of discourse structure rises primarily act on is the question under discussion stack. We can better explain the data by treating rises as telling us something about the current discourse configuration rather than performing an operation on the proffered content of their carrier. If rises signal that a salient question under discussion is unresolved (or similarly a task is unfinished), the post-condition on the discourse is the expectation that someone resolve this question. This may implicate that it is the hearer's responsibility to do this. However, this is not always the case: the speaker may simply be using a rise as a floor holding device, e.g. in listings. What is necessary for the rise to be felicitous, is that the hearer be able to detect what part of the QUD is unresolved.

Looking at rises in an expectations based way helps us to reconcile theoretical approaches with instrumental and impressionistic findings on rises and speaker attitudes. If a response already signals a lack of QUD resolution, e.g. indirect responses or *really*, the rise will mostly serve to heighten the perceived speaker engagement and hence the importance of the proffered contribution to the dialogue. Putting more prosodic effort into a *really*, for example, signals to the hearer that a proposed addition to common ground actually does require some reorganization of the background, i.e. it's surprising. On the other hand, if the semantics and pragmatics of the utterance have a resolving or closing meaning, e.g. a direct answer or a conventionalized parting, adding a rise will require some accommodation. This will in turn be interpreted on an attitudinal level as expressing uncertainty. It seems plausible to say that in task-oriented dialogues task/QUD incompleteness is salient, so rises on task medial moves like instructions and affirmatives are not likely to indicate uncertainty in an epistemic sense of not being able to evaluate a specific proposition. If they indicate some sort of uncertainty it is to do with whether participants are adequately synchronized to complete the task.

Early on, we saw that analyses based on polar and declarative questions tend to

only focus on nuclear rises. While I have argued that rises in fall-rise accents have the same basic meaning as in a proper nuclear rising accent, we still have to take prosodic context into account: the fall in the fall-rises means something too. That is, we do want to treat pitch accents as contributing something separate from the rise. At a first glance, this might indicate that associating meaning with phrase accents, as suggested by Bartels (1999), might it be warranted. However, her specific proposal the L- contributes a [+assert] morpheme is problematic given that declarative questions can fall. Moreover, falling polar questions are quite normal in Southern British English. Once again a mapping between tones and speech act inducing morphemes seems too strong. Instead, it seems that most of the meaning contribution made by pitch accents is relative to their placement, i.e. metrical structure. We can then take falls (i.e. peaks) as the default way to mark metrically strong positions in English. Putting a peak on a syllable is the iconic way of making something prominent pitchwise. Now, we take rising accents (i.e. valleys) to be marked. Combining this with a terminal rise, gives us a conventionalized check meaning. Without it, we basically end up with a delayed peak which we can interpret as a booster signal for contrast. The crucial point is that with adequate contextual interference, the default peak can too be interpreted as questioning. The prediction is that the perception of how much like an actual information seeking question a declarative is will be dependent on the information in the conversational background, as suggested by the data reported in Gunlogson (2008).

From this point of view, final rises signal non-finality, final falls are underspecified, and boundary tones provide separate discourse information to what is provided by pitch accents. However, due to the continuous nature of pitch contours it modelling nuclear tunes in an off-ramp style can give us a clearer indication of what is going on. Direct modelling of the contour gives us a principled way of simultaneous looking at the sign of an intonational shape (e.g. H = concave, L = convex), and the magnitude of the gesture. Using functional decomposition techniques gives us a way of capturing the shape component of ToBI annotations in a more reproducible and less effort intensive way. Using these sorts of continuous measures crucially gives us a better idea of the variance in the data which can help ward off overfitting of meaning to intonational form. Of course, to do the sort of analysis we still have to know what part of the utterance to look at. However, what this really indicates is that the type of annotations we actually need are those to do with structure and attitudinal interpretation, not intonational form.

8.3 Future Directions

Although we have come some way in this dissertation in bridging the gap between theoretical and empirical accounts of prosodic meaning, the overarching question still remains open. The last comment points directly to possible further avenues of research into how prosody affects dialogue interpretation. The studies in this dissertation suggest that to pursue the question of prosodic meaning on a larger scale, we really need some way to efficiently characterize the local context of an utterance. That is its information and metrical structure. A first step towards might be to try to make inferences about response type based on prominence placement using automated techniques for prominence and boundary detection (cf. Rosenberg (2009)). A good data set to start on would, of course, be the IViE corpus for which the non-Cambridge dialogues have not been analyzed. Similarly, looking at perception of rises in different styles for reading-rising and spontaneous-rising dialects should also be fruitful for understanding how style and task expectations affect intonational interpretation.

The dialectal issues are intimately bound up with the question of accommodation. It would be interesting to look at how quickly people accommodate speech from speakers with a different prosodic system. Such studies would have interesting implications for work on speech intelligibility and automated pronunciation scoring for non-native speech. This also suggests further ways to investigate the issues of intonational representation. In particular, further investigation is necessary to get a better understanding of how the height, tilt and convexity maps of shown in this dissertation relate to acoustic perception (i.e. is there are rise or not). Similarly, we would want to see how such a map of acoustic perception relates to linguistic or attitudinal perception (cf. Studdert-Kennedy and Hadding (1973)). Overall, however, to get a truly robust theory of prosodic meaning we need to look at speech representing different task and style conditions, so we can truly understand how the interpretation of prosody varies all the way up the contextual ladder.

Appendix A

Production Experiment Materials

A.1 Short Contexts

A.1.1 Direct Contradictions

(1) A: So, Marianne didn't meet with Lenny
B: No! Marianne did meet with Lenny

(2) A: So, William didn't run away
B: No, William ran away

(3) A: So, Mary didn't remember my birthday
B: No. Mary remembered your birthday

(4) A: So, Emily didn't bring a meringue
B: No. Emily did bring a meringue

(5) A: So, The robbery wasn't really a decoy
B: No. the robbery was really a decoy

(6) A: So, The robbery really wasn't a decoy
B: No. The robbery really was a decoy

(7) A: So, Annemarie isn't in Rio
B: No, Annemarie really is in Rio.

(8) A: So, Annemarie isn't really in Rio
B: No, Annemarie is really in Rio.

A.1.2 Direct Agreements

(9) A: So, Marianne met with Lenny
B: Right. Marianne did meet with Lenny

(10) A: So, William ran away
B: Right. William ran away

(11) A: So, Mary remembered my birthday
B: Right. Mary remembered your birthday

(12) A: So, Emily made a meringue
B: Right. Emily did bring a meringue

(13) A: So, the robbery was really a decoy
B: Right. The robbery was really a decoy

(14) A: So, the robbery really was a decoy
B: Right. The robbery really was a decoy

(15) A: So, Annemarie really is in Rio
B: Right. Annemarie really is in Rio

(16) A: So, Annemarie is really in Rio
B: Right. Annemarie is really in Rio

A.1.3 Indirect Contradictions

(17) A: We haven't got anything connecting Marianne to Lenny
B: Marianne did meet with Lenny

(18) A: There's no reason anyone would suspect the boys
B: William ran away

(19) A: Mary always forgets special occasions
B: Mary remembered your birthday

(20) A: I'm annoyed because nobody brought a dessert!
B: Emily did bring a meringue

(21) A: The boss said that this robbery was Lenny's main job
B: No. The robbery was really a decoy

(22) A: I think the boss is wrong to treat the robbery as a diversion..
B: No. The robbery really was a decoy

(23) A: I heard that Annemarie lied about being overseas
B: No. Annemarie really is in Rio

(24) A: She said she was doing business in New York.
B: No. Annemarie is really in Rio

A.1.4 Indirect Agreements

(25) A: It may be that Marianne is conspiring with Lenny
B: Marianne did meet with Lenny

(26) A: There must be a reason they suspect the boys
B: William ran away

(27) A: The girls are improving, I think
B: Mary remembered your birthday

(28) A: I think there will be something good to eat
B: Emily did bring a meringue

(29) A: It seems that this bank thing was just a diversion
B: Right. The robbery was really a decoy

(30) A: Bill was wrong to say that the robbery wasn't a distraction B: Right. The robbery really was a decoy

(31) A: I think Mary was wrong to say that Annemarie lied about where she is B: Right. Annemarie really is in Rio

(32) A: It seems that Annemarie is not in New York
B: Right. Annemarie is really in Rio

A.1.5 Direct Declarative Questions

(33) A: Marianne met with Lenny
B: What? Marianne did meet with Lenny?

(34) A: William ran away
B: What? William ran away?

(35) A: Mary remembered my birthday
B: What? Mary remembered your birthday?

(36) A: Emily brought a meringue
B: What? Emily did bring a meringue?

(37) A: So, the robbery was a decoy
B: What? The robbery was really a decoy?

(38) A: So, the robbery was a decoy
B: What?. The robbery really was a decoy?

(39) A: So, Annemarie is in Rio
B: What? Annemarie really is in Rio?

(40) A: So, Annemarie is in Rio
B: What? Annemarie is really in Rio?

A.2 Longer Contexts

A.2.1 Discovery/Affirm: Robbery

B: So, what have you found out? What do we have on Marianne's meeting?
A: Well here's some surveillance footage. There's a man and a woman.
A: So, **Marianne did meet with Lenny**
B: As we thought
B: And you think that William was the one who stole the money, not Marianne?
A: **William ran away**
B: From the crime scene?
A: Right. Not very professional, in my opinion.
A: Anyway, we should be getting fingerprint data anytime now.
B: So, what did the others do?
A: **Marianne provided the building schema**, **Lenny provided the password**.
But we don't know who else is involved... Okay, here's the fingerprint data
B: So, **William did the steal money.**
A: Apparently Marianne was also at the scene
B: She isn't usually this sloppy.
B: Maybe we're missing something
A: **Marianne did make some bad mistakes this time.**
A: I'm getting a bad feeling
B: What?
A: **The robbery was really a decoy**
B: I'm thinking the same thing

A: It has been a little too easy
B: But what could be the motive?
A: I have no idea.
B: I've got a message...There's been a big hit downtown
B: It looks like Lenny's gang. And he's left the country.
A: Damn. **The robbery really was a decoy**

A.2.2 Contradictions: Gran's Birthday

B: Gran, Why are you in such a bad mood?
A: I'm not in a bad mood.
B: **You are in a bad mood.**
B: Didn't you have a nice time?
A: Oh I did...even though my own daughter didn't wish me a happy birthday.
B: What? **Annemarie did wish you a happy birthday.**
A: But she didn't come and say it in person, did she?
B: Annemarie's in Rio.
A: She's not really in Rio. I saw her yesterday. She's just avoiding me.
B: **Annemarie really is in Rio.** You talked to her on skype, remember?
A: Well, what about your sister Mary? She's just at home lying in bed.
B: **She did just have surgery**
A: Did she send a card?
B: I'm sure she did.
A: She's forgotten my birthday.
A: Well, she's not as bad as Emily. Going on and on about how much she loves baking but she didn't bring any cake, did she?
B: **Emily did bring a meringue**
A: **Meringue isn't cake.**
B: Well, we had more than enough cake.
A: And nobody even got me so much as a card
B: **Bill got you a cashmere shawl**
B: **Annemarie got you that tree fern**
B: And see! **Mary remembered your birthday**
B: She got you a bottle of port
A: A bottle of port? I hate port.
B: **You did say you liked it when we had it last year**
A: I said no such thing.
A: Anyway, I expect Annemarie will drink it all, after she's tired of pretending she's

overseas.
B: Oh Gran, **Annemarie is really in Rio.**
A: If you say so...

Bibliography

Allen, J. and Core, M. (1997). Draft of DAMSL: Dialog act markup in several layers. *Unpublished manuscript*, 2.

Allwood, J., Nivre, J., and Ahlsen, E. (1992). On the semantics and pragmatics of linguistic feedback. *Journal of semantics*, 9(1):1.

Amir, N., Mixdorff, H., Amir, O., Rochman, D., Diamond, G., Pfitzinger, H., Levi-Isserlish, T., and Abramson, S. (2010). Unresolved anger: Prosodic analysis and classification of speech from a therapeutic setting. In *Speech Prosody 2010*.

Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., et al. (1991). The hcrc map task corpus. *Language and speech*, 34(4):351–366.

Aran, O., Hung, H., and Gatica-Perez, D. (2010). A multimodal corpus for studying dominance in small group conversations. In *Proc. LREC workshop on Multimodal Corpora, Malta*.

Artstein, R. and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Arvaniti, A., Garding, G., et al. (2007). Dialectal variation in the rising accents of american english. *Laboratory phonology*, 9:547–576.

Asher, N. and Lascarides, A. (2003). *Logics of conversation*. Cambridge University Press.

Bales, R. (1969). *Personality and interpersonal behavior*. Holt, Rinehart, and Winston.

Banuazizi, A. and Creswell, C. (1999). Is that a real question? Final rises, final falls and discourse function in yes-no question intonation. In *Papers from the 35th Regional Meeting of the Chicago Linguistic Society*.

Bartels, C. (1999). *The intonation of English statements and questions: a compositional interpretation*. Routledge.

Benotti, L. (2009). Clarification Potential of Instructions. In *Proceedings of the SIGDIAL 2009 Conference*, pages 196–205. Association for Computational Linguistics.

Benus, S., Gravano, A., and Hirschberg, J. (2007). The prosody of backchannels in American English. In *Proceedings of ICPhS 2007*, pages 1065–1068.

Beyssade, C. and Marandin, J. (2007). French intonation and attitude attribution. In *Texas Linguistics Society Conference: Issues at the Semantics-Pragmatics Interface*. Cascadilla Press.

Bolinger, D. (1958). A Theory of Pitch Accent in English. *Word*, 14:109–14.

Bolinger, D. (1972). *Degree words*. Mouton The Hague.

Bolinger, D. (1986). *Intonation and its parts: The melody of language*. Stanford University Press, Stanford, CA.

Bolinger, D. (1989). *Intonation and its uses : melody in grammar and discourse*. Stanford University Press, Stanford, Calif.

Brazil, D. (1997). *The communicative value of intonation in English*. Cambridge University Press.

Britain, D. (1992). Linguistic change in intonation: The use of high rising terminals in new zealand english. *Language Variation and Change*, 4(1):77–104.

Brown, G., Currie, K., and Kenworthy, J. (1980). *Questions of intonation*. Croom Helm London.

Büring, D. (2003). On D-Trees, Beans, and B-Accents. *Linguistics & Philosophy*, 26(5):511–545.

Büring, D. (2008). Whats new and (whats given) in the theory of focus. In *Proceedings of the Berkeley Linguistics Society Meeting*.

Calhoun, S. (2007). *Information structure and the prosodic structure of English: A probabilistic relationship*. PhD thesis, University of Edinburgh.

Calhoun, S. (2010). The centrality of metrical structure in signaling information structure: A probabilistic perspective. *Language*, 86(1):1–42.

Calhoun, S. (2012). The theme/rheme distinction: Accent type or relative prominence? *Journal of Phonetics*, 40(2):329–349.

Calhoun, S., Carletta, J., Brenier, J., Mayo, N., Jurafsky, D., Steedman, M., and Beaver, D. (2010). The NXT-format Switchboard Corpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. *Language resources and evaluation*, pages 1–33.

Campbell, N. (2005). Developments in corpus-based speech synthesis: Approaching natural conversational speech. *IEICE transactions on information and systems*, 88(3):376–383.

Campbell, N. and Beckman, M. (1997). Stress, prominence, and spectral tilt. In *Intonation: Theory, models and applications*.

Campbell, N. and Mokhtari, P. (2003). Voice quality: the 4th prosodic dimension. In *15th ICPhS*, pages 2417–2420.

Carletta, J., Isard, S., Doherty-Sneddon, G., Isard, A., Kowtko, J., and Anderson, A. (1997). The reliability of a dialogue structure coding scheme. *Computational linguistics*, 23(1):13–31.

Carlson, L. (1983). *Dialogue games: An approach to discourse analysis*, volume 17. Springer.

Caspers, J. (2000). Melodic characteristics of backchannels in dutch map task dialogues. In *Sixth International Conference on Spoken Language Processing*.

Charfuelan, M. and Schröder, M. (2011). Investigating the prosody and voice quality of social signals in scenario meetings. *Affective Computing and Intelligent Interaction*, pages 46–56.

Charfuelan, M., Schröder, M., and Steiner, I. (2010). Prosody and voice quality of vocal social signals: the case of dominance in scenario meetings. In *Interspeech'10*.

Cheng, W. and Warren, M. (2005). //CAN i help you //: The use of rise and rise-fall tones in the Hong Kong Corpus of Spoken English. *International Journal of Corpus Linguistics*, 10(1):85–107.

Clark, H. (1996). *Using language*. Cambridge University Press, Cambridge.

Cohen, L. (1992). *An essay on belief and acceptance*. Oxford University Press, USA.

Constant, N. (2007). English Rise-Fall-Rise: A study in the Semantics and Pragmatics of Intonation. Master's thesis, UC Santa Cruz.

Core, M. and Allen, J. (1997). Coding dialogs with the damsl annotation scheme. In *AAAI Fall Symposium on Communicative Action in Humans and Machines*, pages 28–35. Citeseer.

Cruttenden, A. (1981). Falls and rises: meanings and universals. *Journal of Linguistics*, 17(1):77–99.

Cruttenden, A. (1997). *Intonation*. Cambridge Univ Press, Cambridge.

Cruttenden, A. (2007). Intonational diglossia: a case study of glasgow. *Journal of the International Phonetic Association*, 37(03):257–274.

Crystal, D. (1969). *Prosodic Systems and Intonation in English*. Cambridge University Press.

Davis, C., Potts, C., and Speas, M. (2007). The pragmatic values of evidential sentences. In Gibson, M. and Friedman, T., editors, *Proceedings of the 17th Conference on Semantics and Linguistic Theory*, pages 71–88. CLC Publications, Ithaca, NY.

Dilley, L. (2010). Pitch range variation in english tonal contrasts: Continuous or categorical? *Phonetica*, 67(1-2):63–81.

Edlund, J., Heldner, M., and Pelcé, A. (2009). Prosodic features of very short utterances in dialogue. In *Nordic Prosody-Proceedings of the Xth Conference*, pages 56–68. Citeseer.

Eilam, A. (2011). *Explorations in the Informational Component*. PhD thesis, University of Pennsylvania.

Evanini, K. and Lai, C. (2010). The importance of optimal parameter setting for pitch extraction. In *Presented at the 2nd PanAmerican/Iberian Meeting on Acoustics, Cancun, Mexico, 15-19 November 2010*.

Farkas, D. and Bruce, K. (2010). On reacting to assertions and polar questions. *Journal of semantics*, 27(1):81.

Fernandez, R. (2006). *Non-Sentential Utterances in Dialogue: Classification, Resolution and Use*. PhD thesis, Department of Computer Science, Kings College London, University of London.

von Fintel, K. and Gillies, A. (2010). Must... stay... strong. *Natural Language Semantics*. To appear.

Gelman, A. and Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press Cambridge.

Geluykens, R. (1988). On the myth of rising intonation in polar questions. *Journal of Pragmatics*, 12(4):467–485.

Ginzburg, J. (2012). *The Interactive Stance: Meaning for Conversation*. Oxford University Press.

Godfrey, J., Holliman, E., and McDaniel, J. (1992). SWITCHBOARD: telephone speech corpus for research and development. In *ICASSP-92*.

Grabe, E. (2002). Variation adds to prosodic typology. In *Proceedings of the speech prosody 2002 conference*, pages 127–132. Citeseer.

Grabe, E. (2004). Intonational variation in urban dialects of english spoken in the british isles. In Gilles, P. and Peters, J., editors, *Regional variation in intonation*, pages 9–31. Linguistische Arbeiten, Tuebingen, Niemeyer.

Grabe, E., Kochanski, G., and Coleman, J. (2003). Quantitative modelling of intonational variation. *Proc. of SASRTLM*, pages 45–57.

Grabe, E., Kochanski, G., and Coleman, J. (2005). The intonation of native accent varieties in the british isles: Potential for miscommunication. In Dziubalska-Kolaczyk, K. and Przedlacka, J., editors, *English pronunciation models: A changing scene*, volume 21 of *Linguistic Insights Series*. Peter Lang Publishing, Frankfurt am Main.

Grabe, E., Kochanski, G., and Coleman, J. (2007). Connecting intonation labels to mathematical descriptions of fundamental frequency. *Language and Speech*, 50(3):281–310.

Grabe, E. and Post, B. (2002). Intonational Variation in the British Isles. In *Proceedings of the Speech Prosody 2002 Conference*, pages 343–346.

Grabe, E., Post, B., and Nolan, F. (2001). Modelling intonational Variation in English. The IViE system. In Puppel, S. and Demenko, G., editors, *Proceedings of Prosody 2000, Adam Mickiewitz University, Poznan, Poland*, pages 51–57.

Grabe, E., Post, B., Nolan, F., and Farrar, K. (2000). Pitch accent realization in four varieties of british english. *Journal of Phonetics*, 28(2):161–185.

Gravano, A. (2009). *Turn-Taking and Affirmative Cue Words in Task-Oriented Dialogue*. PhD thesis, Columbia University.

Gravano, A., Benus, S., Chavez, H., Hirschberg, J., and Wilcox, L. (2007). On the role of context and prosody in the interpretation of okay. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 800–807. Association for Computational Linguistics.

Gravano, A., Benus, S., Hirschberg, J., German, E. S., and Ward, G. (2008). The effect of prosody and semantic modality on the assessment of speaker certainty. In *Proceedings of 4th Speech Prosody Conference, Campinas, Brazil*.

Gravano, A. and Hirschberg, J. (2009). Turn-Yielding Cues in Task-Oriented Dialogue. In *Proceedings of SIGDIAL 2009*, pages 253–261.

Grice, H. (1975). Logic and conversation. *Syntax and semantics*, 3:41–58.

Groenendijk, J. and Stokhof, M. (1984). *Studies on the semantics of questions and the pragmatics of answers*. PhD thesis, Universiteit van Amsterdam.

Grosz, B. and Sidner, C. (1986). Attention, intentions, and the structure of discourse. *Computational linguistics*, 12(3):175–204.

Gubian, M., Cangemi, F., and Boves, L. (2010). Automatic and data driven pitch contour manipulation with functional data analysis. In *Proceedings of Speech Prosody 2010, Chicago, USA*.

Gunlogson, C. (2003). *True to form: Rising and falling declaratives as questions in English*. Routledge.

Gunlogson, C. (2008). A question of commitment. *Belgian Journal of Linguistics*, 22(1):101–136.

Gussenhoven, C. (1984). *On the grammar and semantics of sentence accents*. Walter de Gruyter.

Gussenhoven, C. (2002). Intonation and Interpretation: Phonetics and Phonology. In *Speech Prosody 2002, International Conference*. ISCA.

Gussenhoven, C. (2004). *The phonology of tone and intonation*. Cambridge University Press.

Gussenhoven, C. and Chen, A. (2000a). Universal and language-specific effects in the perception of question intonation. In *Sixth International Conference on Spoken Language Processing*.

Gussenhoven, C. and Chen, A. (2000b). Universal and Language-Specific Effects in the Perception of Question Intonation. In *Sixth International Conference on Spoken Language Processing*. ISCA.

Gussenhoven, C. and Rietveld, T. (2000). The Behavior of H and L Under Variations in Pitch Range in Dutch Rising Contours. *Language and Speech*, 43(2):183–203.

Gutzmann, D. and Castroviejo Miró, E. (2009). The Dimensions of VERUM. In *Colloque de Syntaxe et Sḿantique à Paris*.

Guy, G., Horvath, B., Vonwiller, J., Daisley, E., and Rogers, I. (1986). An intonational change in progress in australian english. *Language in Society*, 15(1):23–51.

Haan, J. (2002). *Speaking of Questions: An Exploration of Dutch Question Intonation*. PhD thesis, Utrecht: LOT Graduate School of Linguistics.

Hacquard, V. (2010). On the event relativity of modal auxiliaries. *Natural language semantics*, 18(1):79–114.

Halliday, M. (1970). *A Course in spoken english: intonation*. Oxford University Press, Oxford.

Heim, I. (1983). File change semantics and the familiarity theory of definiteness. *Formal Semantics*, pages 223–248.

Heldner, M., Edlund, J., Laskowski, K., and Pelcé, A. (2008). Prosodic features in the vicinity of silences and overlaps. In *Proc. 10th Nordic Conference on Prosody*, pages 95–105. Citeseer.

Hirschberg, J. (2000). A corpus-based approach to the study of speaking style. *Prosody: Theory and experiment*, pages 335–350.

Hirschberg, J. (2002a). Communication and prosody: Functional aspects of prosody. *Speech Communication*, 36(1-2):31–43.

Hirschberg, J. (2002b). The pragmatics of intonational meaning. In *Speech Prosody 2002, International Conference*.

Hirschberg, J. and Litman, D. (1994). Empirical studies on the disambiguation of cue phrases. *Computational linguistics*, 19(3):501–530.

Hirschberg, J. and Ward, G. (1992). The influence of pitch range, duration, amplitude and spectral features on the interpretation of the rise-fall-rise intonation contour in English. *Journal of phonetics*.

Hobbs, J. (1990). The pierrehumbert-hirschberg theory of intonational meaning made simple: Comments on pierrehumbert and hirschberg. *Intentions in Communication*, pages 313–323.

Hockey, B. (1993). Prosody and the role of okay and uh-huh in discourse. In *Proceedings of the eastern states conference on linguistics*, pages 128–136. Citeseer.

Höhle, T. (1992). Über Verum-Fokus im Deutschen. *Informationsstruktur und Grammatik*, 4:112–142.

Hyman, L. and Watters, J. (1984). Auxiliary Focus. *Studies in African linguistics*, 15(3):233–273.

Jackendoff, R. (1974). *Semantic Interpretation in Generative Grammar*. MIT Press.

Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., et al. (2003). The icsi meeting corpus. In *Proceedings of ICASSP'03*. IEEE.

Jarman, E. and Cruttenden, A. (1976). Belfast intonation and the myth of the fall. *Journal of the International Phonetic Association*, 6(01):4–12.

Jayagopi, D., Hung, H., Yeo, C., and Gatica-Perez, D. (2009). Modeling dominance in group conversations using nonverbal activity cues. *Audio, Speech, and Language Processing, IEEE Transactions on*, 17(3):501–513.

Jurafsky, D., Shriberg, E., and Biasca, D. (1997). Switchboard-DAMS Labeling Project Coders Manual. Technical Report 97-02, University of Colorado Institute of Cognitive Science.

Kadmon, N. (2001). *Formal pragmatics*. Blackwell.

Kennedy, C. (1999). *Projecting the adjective: The syntax and semantics of gradability and comparison*. Garland Press, New York.

Kennedy, C. and McNally, L. (2005). Scale structure, degree modification, and the semantics of gradable predicates. *Language*, 81(2):345–381.

Kennedy, C. and McNally, L. (2010). Color, context, and compositionality. *Synthese*, 174(1):79–98.

Kingdon, R. (1958). *The groundwork of English intonation*. Longmans.

Knowles, G. (1978). The nature of phonological variables in scouse. *Sociolinguistic patterns in British English*, pages 80–90.

Kochanski, G., Grabe, E., Coleman, J., and Rosner, B. (2005). Loudness predicts prominence: Fundamental frequency lends little. *The Journal of the Acoustical Society of America*, 118:1038.

Kowtko, J. (1997). *The function of intonation in task-oriented dialogue*. PhD thesis, University of Edinburgh.

Kowtko, J., Isard, S., and Doherty, G. (1991). Conversational games within dialogue. In *Proceedings of the ESPRIT Workshop on Discourse Coherence*.

Kratzer, A. (1981). The notional category of modality. In Eikmeyer, H. and Rieser, H., editors, *Words, worlds, and contexts. New approaches in word semantics*, pages 38–74. Mouton de Gruyter, New York.

Kratzer, A. (2004). Interpreting focus: Presupposed or expressive meanings? A comment on Geurts and van der Sandt. *Theoretical Linguistics*, 30(1):123–136.

Krifka, M. (2001). For a structured meaning account of questions and answers. *Audiatur Vox Sapientia. A Festschrift for Arnim von Stechow*, 52:287–319.

Krippendorff, K. (2004). Reliability Content Analysis: Some Common Misconceptions and Recommendations. *Human Communication Research*, 30(3):411–433.

Ladd, D. (2008). *Intonational phonology*, volume 119. Cambridge Univ Pr.

Ladd, D., Silverman, K., Tolkmitt, F., Bergmann, G., and Scherer, K. (1985). Evidence for the independent function of intonation contour type, voice quality, and f0 range in signaling speaker affect. *Journal of the Acoustical Society of America*, 78(2):435–444.

Ladd, R. D. (1981). A first look at the semantics and pragmatics of negative questions and tag questions. *Proceedings from the Annual Meeting of the Chicago Linguistic Society*, 17:164–171.

Levow, G., Duncan, S., and King, E. (2010). Cross-cultural investigation of prosody in verbal feedback in interactional rapport. In *Eleventh Annual Conference of the International Speech Communication Association*.

Liberman, M. (1979). *The intonational system of English*. Garland, New York, NY.

Liberman, M. and Prince, A. (1977). On stress and linguistic rhythm. *Linguistic inquiry*, 8(2):249–336.

Liberman, M. and Sag, I. (1974). Prosodic form and discourse function. In *Tenth Regional Meeting, Chicago Linguistic Society*, pages 416–427.

Liscombe, J. (2007). *Prosody and Speaker State: Paralinguistics, Pragmatics, and Proficiency*. PhD thesis, Columbia University.

Litman, D., Rotaru, M., and Nicholas, G. (2009). Classifying Turn-Level Uncertainty Using Word-Level Prosody. In *Proceedings of Interspeech'09*.

Liu, F. and Xu, Y. (2005). Parallel encoding of focus and interrogative meaning in mandarin intonation. *Phonetica*, 62(2-4):70–87.

Local, J., Kelly, J., and Wells, W. (1986). Towards a phonology of conversation: turn-taking in tyneside english. *Journal of Linguistics*, 22(2):411–437.

Mayo, C., Aylett, M., and Ladd, D. (1997). Prosodic transcription of glasgow english: An evaluation study of glatobi. In *Intonation: Theory, Models and Applications*.

McLemore, C. (1991). *The pragmatic interpretation of English intonation: Sorority speech*. PhD thesis, University of Texas at Austin.

Merin, A. and Bartels, C. (1997). Decision-Theoretic Semantics for Intonation. Technical Report Bericht nr. 88., Universitt Stuttgart and Universitt Tübingen.

Meteer, M. and Taylor, A. (1995). Dysfluency annotation stylebook for the switchboard corpus.

Nilsenova, M. (2006). *Rises and Falls. Studies in the semantics and pragmatics of intonation*. PhD thesis, University of Amsterdam.

O'Connor, J. and Arnold, G. (1961). *Intonation of colloquial English: a practical handbook*. Longmans.

Ohala, J. (1984). An ethological perspective on common cross-language utilization of F0 of voice. *Phonetica*, 41(1):1–16.

Ostendorf, M., Shafran, I., Shattuck-Hufnagel, S., Carmichael, L., and Byrne, W. (2001). A prosodically labeled database of spontaneous speech. In *ISCA Tutorial and Research Workshop (ITRW) on Prosody in Speech Recognition and Understanding*.

Palmer, H. (1922). *English intonation with systematic exercises*. W. Heffer & Sons.

Pierrehumbert, J. (1980). *The phonology and phonetics of English intonation*. PhD thesis, MIT.

Pierrehumbert, J. and Beckman, M. (1988). *Japanese Tone Structure*. The MIT Press.

Pierrehumbert, J. and Hirschberg, J. (1990). The meaning of intonational contours in the interpretation of discourse. In Cohen, P., Morgen, J., and Pollack, M., editors, *Intentions in Communication*. MIT Press, Cambridge.

Pon-Barry, H. (2008). Prosodic manifestations of confidence and uncertainty in spoken language. In *Proceedings of Interspeech'08*.

Portner, P. (2004). The Semantics of Imperatives within a Theory of Clause Types. In Watanabe, K. and Young, R. B., editors, *Proceedings of Semantics and Linguistic Theory 14*, pages 235–252. CLC Publications, Ithaca, NY.

Portner, P. (2007). Instructions for Interpretation as Separate Performatives. In Schwabe, K. and Winkler, S., editors, *On Information Structure, Meaning and Form*, pages 407–426. John Benjamins.

Potts, C. (2005). *The Logic of Conventional Implicatures*. Oxford Studies in Theoretical Linguistics. Oxford University Press, Oxford.

Potts, C. (2007). The expressive dimension. *Theoretical Linguistics*, 33(2):165–198.

Prasad, S. and Bali, K. (2010). Prosody cues for classification of the discourse particle" hã" in hindi. In *Eleventh Annual Conference of the International Speech Communication Association*.

Purver, M. (2004). *The Theory and Use of Clarification Requests in Dialogue*. PhD thesis, King's College, University of London.

Ramsay, J. (2006). *Functional data analysis*. Wiley Online Library.

Rangarajan Sridhar, V., Bangalore, S., and Narayanan, S. (2009). Combining lexical, syntactic and prosodic cues for improved online dialog act tagging. *Computer Speech & Language*, 23(4):407–422.

Raux, A., Langner, B., Bohus, D., Black, A., and Eskenazi, M. (2005). Let's go public! taking a spoken dialog system to the real world. In *Ninth European Conference on Speech Communication and Technology*.

Reese, B. (2007). *Bias in Questions*. PhD thesis, University of Texas at Austin.

Renals, S., Hain, T., and Bourlard, H. (2007). Recognition and understanding of meetings the ami and amida projects. In *Automatic Speech Recognition & Understanding, 2007. ASRU. IEEE Workshop on*, pages 238–247. IEEE.

Rietveld, T., Gussenhoven, C., Wichmann, A., and Grabe, E. (1999). The Communicative Effects of Rising and Falling Pitch Accents in British English and Dutch. In *ESCA Tutorial and Research Workshop (ETRW) on Dialogue and Prosody*. ISCA.

Roberts, C. (1996). Information structure in discourse: Towards an integrated formal theory of pragmatics. *Working Papers in Linguistics-Ohio State University Department of Linguistics*, pages 91–136.

Romero, M. (2006). Biased Yes/No Questions: The Role of VERUM. *SPRACHE UND DATENVERARBEITUNG*, 30(1):9.

Romero, M. and Han, C. (2004). On negative yes/no questions. *Linguistics and Philosophy*, 27(5):609–658.

Rooth, M. (1985). *Association with focus*. University of Massachusetts.

Rosenberg, A. (2009). *Automatic Detection and Classification of Prosodic Events*. PhD thesis, Columbia University.

Saget, S., Guyomard, M., et al. (2006). Goal-oriented dialog as a collaborative subordinated activity involving collective acceptance. In *Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue (Brandial 2006), University of Potsdam, Potsdam, Germany*, page 131.

Salamin, H., Vinciarelli, A., Truong, K., and Mohammadi, G. (2010). Automatic role recognition based on conversational and prosodic behaviour. In *Proceedings of the international conference on Multimedia*, pages 847–850. ACM.

Savino, M. (2011). The intonation of backchannels in italian task-oriented dialogues: cues to turn-taking dynamics, information status and speakers attitude. In *5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*.

Scherer, K., Ladd, D., and Silverman, K. (1984). Vocal cues to speaker affect: Testing two models. *Journal of the Acoustical Society of America*, 76(5):1346–1356.

Schiffrin, D. (1994). *Discourse markers*. Cambridge University Press.

Schlangen, D. (2004). Causes and Strategies for Requesting Clarification in Dialogue. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue (SIGDIAL 04), Boston, USA*.

Searle, J. (1969). *Speech acts: An essay in the philosophy of language*. Cambridge Univ Pr.

Selkirk, E. (1986). *Phonology and Syntax: The Relation Between Sound and Structure*. Current Studies in Linguistics. MIT Press.

Selkirk, E. (1995). Sentence prosody: Intonation, stress, and phrasing. In *The handbook of phonological theory*, pages 550–569. Blackwell, London.

Shattuck-Hufnagel, S. and Turk, A. (1996). A prosody tutorial for investigators of auditory sentence processing. *Journal of psycholinguistic research*, 25(2):193–247.

Shimojima, A., Katagiri, Y., Koiso, H., and Swerts, M. (2002). Informational and dialogue-coordinating functions of prosodic features of Japanese echoic responses. *Speech Communication*, 36(1-2):113–132.

Shriberg, E., Bates, R., Taylor, P., Stolcke, A., Jurafsky, D., Ries, K., Coccaro, N., Martin, R., and Meteer, M. (1998). Can Prosody Aid the Automatic Classification of Dialog Acts in Conversational Speech? *Language and Speech*, 41(3-4):443–492.

Shriberg, E. and Stolcke, A. (2004). Direct modeling of prosody: An overview of applications in automatic speech processing. In *Speech Prosody 2004*. Citeseer.

Shriberg, E., Stolcke, A., Hakkani-Tur, D., and Tur, G. (2000). Prosody-based automatic segmentation of speech into sentences and topics. *Speech communication*, 32(1-2):127–154.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., and Hirschberg, J. (1992). ToBI: A standard for labeling English prosody. In *Second International Conference on Spoken Language Processing*, volume 2, pages 867–870.

Sluijter, A. and Van Heuven, V. (1996). Spectral balance as an acoustic correlate of linguistic stress. *JASA*, 100(4):2471–2485.

Stalnaker, R. (1978). Assertion. *Formal Semantics*, pages 147–161.

Stalnaker, R. (2002). Common ground. *Linguistics and Philosophy*, 25(5):701–721.

Steedman, M. (2000). Information Structure and the Syntax-Phonology Interface. *Linguistic Inquiry*, 31(4):649–689.

Steedman, M. (2007). Information-structural semantics for English intonation. In Lee, C., Gordon, M., and Büring, D., editors, *Topic and Focus*, pages 245–264. Springer.

Strassel, S. (2003). Simple Metadata Annotation Specification V5.0. *Linguistic Data Consortium, Philadelphia*.

Studdert-Kennedy, M. and Hadding, K. (1973). Auditory and linguistic processes in the perception of intonation contours. *Language and Speech*, 16(4):293.

Swerts, M. and Geluykens, R. (1994). Prosody as a marker of information flow in spoken discourse. *Language and speech*, 37(1):21–43.

Syrdal, A. and McGory, J. (2000). Inter-transcriber reliability of tobi prosodic labeling. In *Sixth International Conference on Spoken Language Processing*.

Taylor, P., King, S., Isard, S., and Wright, H. (1998). Intonation and dialog context as constraints for speech recognition. *Language and Speech*, 41(3-4):493–512.

Truong, K. and Heylen, D. (2010). Disambiguating the functions of conversational sounds with prosody: the case of'yeah'. In *Proceedings of Interspeech 2010*. International Speech Communication Association (ISCA).

Tür, G., Hakkani-Tür, D., Stolcke, A., and Shriberg, E. (2001). Integrating prosodic and lexical cues for automatic topic segmentation. *Computational linguistics*, 27(1):31–57.

Uldall, E. (1962). Ambiguity: Question or statement? or'are you asking me or telling me?'. In *Proceedings of the Fourth International Congress of Phonetic Sciences*, pages 770–83.

Vallduví, E. (1990). *The Informational Component*. PhD thesis, University of Pennsylvania.

Veltman, F. (1996). Defaults in update semantics. *Journal of philosophical logic*, 25(3):221–261.

Vinciarelli, A., Valente, F., Yella, S., and Sapru, A. (2011). Understanding social signals in multi-party conversations: Automatic recognition of socio-emotional roles in the ami meeting corpus. In *2011 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 374–379. IEEE.

Von Stechow, A. (1991). Focusing and backgrounding operators. In *Discourse Particles: Pragmatics & Beyond*, pages 37–84. John Benjamins, Amsterdam.

Wagner, M. (2010). Contrastive topics decomposed. Unpublished manuscript.

Ward, G. and Hirschberg, J. (1985). Implicating uncertainty: The pragmatics of fall-rise intonation. *Language*, 61(4):747–776.

Ward, G. and Hirschberg, J. (1988). Intonation and propositional attitude: the pragmatics of L*+ HLH%. In *Proceedings of the Fifth Eastern States Conference on Linguistics*, pages 512–522.

Ward, N. (2004). Pragmatic functions of prosodic features in non-lexical utterances. In *Speech Prosody 2004*, volume 4, pages 325–328.

Ward, N. and Tsukahara, W. (2000). Prosodic features which cue back-channel responses in English and Japanese. *Journal of Pragmatics*, 32(8):1177–1207.

Ward, N. G. and Escalante-Ruiz, R. (2009). Using Subtle Prosodic Variation to Acknowledge the User's Current State. In *Proceedings of Interspeech'09*.

Wilks, Y., Catizone, R., Worgan, S., and Turunen, M. (2011). Review: Some background on dialogue management and conversational speech for dialogue systems. *Computer Speech and Language*, 25(2):128–139.

Wrede, B. and Shriberg, E. (2003). Spotting "hot spots" in meetings: Human judgments and prosodic cues. In *Eighth European Conference on Speech Communication and Technology*.

Wright Hastie, H., Poesio, M., and Isard, S. (2002). Automatically predicting dialogue structure using prosodic features. *Speech Communication*, 36(1-2):63–79.

Xu, Y. (2005-2011). ProsodyPro.praat. http://www.phon.ucl.ac.uk/home/yi/ProsodyPro/.

Yuan, J. and Liberman, M. (2008). Speaker identification on the SCOTUS corpus. *Journal of the Acoustical Society of America*, 123(5):3878.

Yuan, J., Shih, C., and Kochanski, G. (2002). Comparison of declarative and interrogative intonation in Chinese. In *Speech Prosody 2002, International Conference*. ISCA.

Zaroukian, E. (2011). Rising intonation and uncertainty. In *LSA Annual Meeting Extended Abstracts*.